ARI Research Note 91-28

Project A:
Improving the Selection, Classification, and Utilization of Army Enlisted Personnel

Literature Review: Cognitive Abilities—Theory, History, and Validity

Jody L. Toquam, VyVy A. Corpe, and Marvin D. Dunnette
Personnel Decisions Research Institute

Selection and Classification Technical Area
Michael G. Rumsey, Chief

Manpower and Personnel Policy Research Laboratory
Zita M. Simutis, Director

February 1991

United States Army
Research Institute for the Behavioral and Social Sciences

UNCLASSIFIED

ITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

EPORT SECURITY CLASSIFICATION lassified	1b. RESTRICTIVE MARKINGS --
ECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
ECLASSIFICATION/DOWNGRADING SCHEDULE	
RFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Research Note 91-28
AME OF PERFORMING ORGANIZATION an Resources Research anization; 6b. OFFICE SYMBOL (If applicable) --	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute
DDRESS (City, State, and ZIP Code) 0 South Washington Street xandria, VA 22314-4499	7b. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600
IAME OF FUNDING/SPONSORING RGANIZATION U.S. Army Research titute for the Behavioral Social Sciences; 8b. OFFICE SYMBOL (If applicable) PERI-R	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903-82-C-0531
DDRESS (City, State, and ZIP Code) 1 Eisenhower Avenue xandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. 63731A; PROJECT NO. 792; TASK NO. 232; WORK UNIT ACCESSION NO. C1

ITLE (Include Security Classification)

erature Review: Cognitive Abilities--Theory, History, and Validity

ERSONAL AUTHOR(S)

uam, Jody L.; Corpe, VyVy A.; and Dunnette, Marvin D.

TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
erim	FROM 82/01 TO 83/12	1991, February	271

UPPLEMENTARY NOTATION Personnel Decisions Research Institute is a subcontractor of HumRRO. pared under Project A: Improving the Selection, Classification, and Utilization of Army isted Personnel.

COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)	
FIELD	GROUP	SUB-GROUP	Abilities	Criterion measures
			Army-wide measures	Literature review
			Cognitive abilities	Personnel classification (Continued)

BSTRACT (Continue on reverse if necessary and identify by block number)

The research described in this report was performed under Project A, the Army's large-ale, multiyear manpower and personnel effort for improving the selection, classification, l utilization of Army personnel. This report is one of three derived from an extensive erature review to identify constructs that might enhance the accuracy of the Army screen- system (the other two reports deal, respectively, with psychomotor abilities and with lity of temperament, biodata, and interests assessment). Researchers identified and de-ied the constructs included in the cognitive ability domain and summarized early efforts measure intelligence and contemporary cognitive taxonomies and testing techniques. They scribed the development and use of cognitive ability measures for vocational guidance and diction of work performance and discussed social and legal implications of such measure-it. They summarized the validity evidence for traditional psychometric measures. On the sis of the cognitive abilities assessed in four widely used test batteries and two lines extensive research, a cognitive taxonomy that contains nine ability (Continued)

ISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
NAME OF RESPONSIBLE INDIVIDUAL chael G. Rumsey	22b. TELEPHONE (Include Area Code) (202) 274-8275; 22c. OFFICE SYMBOL PERI-RS

orm 1473, JUN 86 Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

ARI Research Note 91-28

18. SUBJECT TERMS (Continued)

Personnel selection
Predictor measures
Project A

19. ABSTRACT (Continued)

factors is suggested. These factors are verbal, number facility, spatial abilities, reasoning, memory, fluency, perception, perceptual speed and accuracy, and mechanical aptitude.

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

JON W. BLADES
COL, IN
Commanding

Research accomplished under contract for the Department of the Army

Human Resources Research Organization

Technical review by

Mark Czarnolewski
Deirdre Knapp

Accession For		
NTIS GRA&I		☒
DTIC TAB		☐
Unannounced		☐
Justification		
By		
Distribution/		
Availability Codes		
Dist	Avail and/or Special	
A-1		

QUALITY INSPECTED

NOTICES

DISTRIBUTION: This report has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The views, opinions, and findings in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other authorized documents.

LITERATURE REVIEW: COGNITIVE ABILITIES--THEORY, HISTORY, AND VALIDITY

PREFACE

This Research Note is one of three that present the results of a literature review conducted as part of Project A, a large-scale, multiyear research program intended to improve the selection and classification system for initial assignment of persons to U.S. Army Military Occupational Specialties. The research is sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI).

The three Research Notes cover measures of human abilities, interests, and other attributes. They are

Literature Review: Cognitive Abilities--Theory, History, and Validity by Jody L. Toquam, VyVy A. Corpe, and Marvin D. Dunnette. ARI Research Note 91-28.

Literature Review: Validity and Potential Usefulness of Psychomotor Ability Tests for Personnel Selection and Classification by Jeffrey J. McHenry and Sharon R. Rose. ARI Research Note 88-13. (AD A193 558)

Literature Review: Utility of Temperament, Biodata, and Interest Assessment for Predicting Job Performance by Leaetta M. Hough, Editor. ARI Research Note 88-02. (AD A192 109)

The findings presented in these documents were used to develop a battery of new tests and inventories for use in Project A. The focus of that development effort was to identify abilities and other human attributes that seemed "best bets" for predicting soldiers' job performance, and then to develop new measures for those attributes. These Research Notes, however, have usefulness beyond that particular applied problem. Many issues pertinent to the measurement and use of human abilities are described and discussed in each of them.

The Research Notes describe the results and findings of the literature review, but do not describe the literature search process itself. Therefore, we provide a description of that process here.

The literature search was conducted by three research teams from the Personnel Decisions Research Institute. Each team was responsible for one of the three areas of human abilities or characteristics that are reported in the Research Notes: cognitive abilities; psychomotor abilities; and noncognitive characteristics, such as vocational interests, biographical data, and measures of temperament. While these domains were convenient for purposes of organizing and conducting literature search activities, they were not used as (nor intended to be) a final taxonomy of possible predictor measures.

The major part of the literature search was conducted in late 1982 and early 1983. Within each of the three areas, the teams carried out essentially the same steps:

1. Compile an exhaustive list of reports, articles, books, or other sources that were possibly relevant to Project A.

2. Review each item and determine its relevance to the project by examining the title and abstract (or other brief review).

3. Obtain the relevant sources identified in the second step.

4. For relevant materials, conduct a thorough review and transfer applicable information onto special review forms developed for the project.

In the first step, several activities were designed to ensure that the list would be as comprehensive as possible. Several computerized searches of relevant data bases were performed. Across all three ability areas, more than 10,000 potential sources were identified using the computer searches (Many of these sources were identified as relevant in more than one area and were counted more than once.)

In addition to the computerized searches, reference lists were obtained from recognized experts in each area, emphasizing the most recent research in the field. Several annotated bibliographies were obtained from military research laboratories. Finally, the last several years' editions of research journals frequently used in each ability area were scanned, as were more general sources such as textbooks, handbooks, and appropriate chapters in the Annual Review of Psychology (which reviews the most recent research in a number of conceptually distinct areas of psychology).

The majority of the items identified in the first steps proved irrelevant to the applied purpose--the identification and development of promising measures for personnel selection in the U.S. Army. These irrelevant sources were weeded out in Step 2.

The relevant sources were obtained and reviewed, and team members completed two forms for each source: an Article Review form and a Predictor Review form (several of the latter could be prepared for each source). These forms were designed to capture, in a standard format, the essential information about the reviewed sources.

The Article Review form contained seven sections: citation, abstract, list of predictors (keyed to the Predictor Review forms), description of criterion measures, description of sample(s), description of methodology, other results, and reviewer's comments. The Predictor Review form also contained seven sections: description of predictor, reliability, norms/descriptive statistics, correlations with other predictors, correlations with criteria, adverse impact/differential validity/test fairness, and reviewer's recommendations (about the usefulness of the predictor). Each predictor was tentatively classified into an initial working taxonomy of predictor constructs.

The Review forms and the actual sources that had been located were used in two primary ways for Project A purposes. First, three working documents were written, one for each of the three areas. These working documents later evolved into the three Research Notes named above. These documents identified and summarized the literature important to the research being conducted, the most appropriate organization or taxonomy of the constructs in each area, and the validities of the various measures for different types of job performance criteria. Second, the predictors identified in the review were subjected to further, structured scrutiny in order to select tests and inventories for use in later activities of Project A.

As a set, the three Research Notes should provide a valuable resource for scientists, researchers, and personnel practitioners interested in the measurement of individual differences in humans for various applied purposes, but especially for selection and classification.

LITERATURE REVIEW: COGNITIVE ABILITIES--THEORY, HISTORY, AND VALIDITY

AUTHOR'S PREFACE

The literature summarized in this document represents the first major research activity undertaken as part of Task 2 (Experimental Measure Development) in Project A. We began the literature search and review in 1982; the draft literature review document was completed in 1985. Thus, the bulk of the literature cited in this review is from sources available in 1985 and earlier.

Our objective in preparing this review was to present a discussion of the salient topics and major issues related to cognitive ability measurement. These topics have continued to be issues for research in the years since we completed a draft of this review. We have had the opportunity to revise the review, most recently in Summer 1989. While completing this latest revision, we attempted to clarify and update certain topic areas with more recent citations. We feel that the information provided in this review provides the reader with ample information for evaluating the major issues in cognitive ability test development and for understanding our perspective in recommending measures to supplement the Armed Services Vocational Aptitude Battery (ASVAB).

LITERATURE REVIEW: COGNITIVE ABILITIES--THEORY, HISTORY, AND VALIDITY

CONTENTS

LIST OF TABLES

LIST OF FIGURES

SECTION I

LITERATURE REVIEW: COGNITIVE ABILITIES -- THEORY, HISTORY, AND VALIDITY

The domain of cognitive and perceptual abilities includes mental and sensory processes by which verbal, spatial, numerical, and figural information is acquired, retained, manipulated, integrated, and reconfigured[1]. The above statement serves as only a crude definition of the many possible types of mental processes that can be measured within the cognitive ability domain. This is because several existing theories suggest that the number of cognitive abilities vary from as few as 1 to more than 100 distinct cognitive abilities. Thus, one major task for this literature review is to identify the number and types of cognitive ability constructs that may be used to predict training and performance outcomes in the U.S. Military Services. In particular, we are interested in identifying cognitive abilities that may be used to enhance the accuracy of the current U.S. Army screening system.

Theoretically and historically, the measurement of distinct cognitive abilities has its roots in a single construct, intelligence. Definitions of intelligence vary according to focus. For example, intelligence can be defined in terms of routine (day-to-day) behaviors, general mental processes, psychometric characteristics, and societal demands. Examples of each of these are discussed in turn.

Within our society, individuals form ideas about the routine or day-to-day behaviors that constitute intelligence. Thus, conventional wisdom (from non-psychologists) is one source for a definition of intelligence. Sternberg, Conway, Ketron, and Bernstein (1981) learned from lay persons that intelligence includes characteristics such as problem-solving, verbal ability, social competence, character, and interest in culture and learning. According to conventional wisdom, then, intelligence may be observed in a wide variety of human behaviors and in many diverse situations (e.g., observed in social settings and self-reported interests).

In terms of general mental processes, several distinguished researchers investigating intelligence attempted to provide a definition of the construct. Thorndike and his colleagues (1921) summarized these definitions:

the power of good responses (Thorndike)
the ability to carry on abstract thinking (Terman)
the ability to learn (Buckingham)
the capacity to acquire capacity (Woodrow).

[1]For the remainder of this report, the term <u>cognitive abilities</u> is used to refer to both cognitive and perceptual ability constructs.

These definitions were intended to describe what tests of intelligence are designed to measure. Although these definitions appear very similar, Thorndike indicates that this group of distinguished researchers failed to reach a consensus in defining intelligence.

A third means of defining intelligence is to view it from a measurement and psychometric perspective. At a very simple level, intelligence has been defined as "that which a properly standardized test measures" (Atkinson, Atkinson & Hilgard, 1983). In terms of complex psychometrics, intelligence has been defined as the first general factor obtained from a factor analysis of correlations among several different types of mental ability tests. In the latter case, intelligence is termed g, because it represents the general factor underlying all intelligence and cognitive ability tests.

In terms of societal demands, Anastasi (1983) has examined the nature of the construct of intelligence as it evolved through years of research in developmental, cross-cultural, learning, and cognitive psychology. In her summary, Anastasi concludes that intelligence is composed of several traits and these traits and the level of their development reflect, in part, a person's age and the demands and reinforcements in the environment. The composition of intelligence, then, may vary with age, level of development, and cultural context. Further, within a particular culture, several factors may be associated with this variation. These include opportunity to develop different cognitive skills and to accumulate different knowledges, as well as motivational and attitudinal factors. In general, Anastasi states that traditional intelligence tests measure "a cluster of abilities or traits demanded in modern technologically advanced societies" (p. 182). Within our culture, the trait or ability cluster is developed by formal schooling and may, therefore, be considered "academic intelligence."

Definitions of intelligence from the above perspectives are used to guide the development of a taxonomy of cognitive abilities. That is, in this review we attempt to identify a cognitive ability taxonomy that reflects distinct mental processes. Thus, unlike the definitions of intelligence provided by the panel of experts that Thorndike and his colleagues (1921) polled, we expect to isolate several cognitive ability constructs in which the specific, rather than general, mental processes measured by these constructs are identified.

As a means to define these ability constructs, this review places strong emphasis on psychometric evidence supporting measurement adequacy and for isolating distinct cognitive abilities. Further, we concur with Anastasi's view of intelligence (and cognitive abilities) that is driven in large part by societal demands. In the context of our review (i.e., identifying cognitive abilities likely to enhance job performance in the U.S. Army), we view job requirements as the proxy for societal demands.

In terms of day-to-day definitions of intelligence, such as that provided by Sternberg et al. (1981), it is important to recognize that intelligence and cognitive abilities influence routine, day-to-day behaviors. In terms of measurement, however, definitions of intelligence and cognitive abilities will be limited to tests measuring mental and sensory processes, thereby avoiding measures that overlap with temperament and biographical constructs.

OVERVIEW OF REPORT

The major purpose of this review is to identify a taxonomy of cognitive ability constructs that have meaning in terms of psychometric characteristics and in terms of external demands (i.e., job requirements). The rationale followed to identify these constructs is described in the following sections. First, however, we provide below an overview of the contents of this report.

The next section provides a historical perspective of intelligence and cognitive abilities measurement from very early times to the present. Also included in this section, is a discussion of the different theories of intelligence; these theories lead to the development of cognitive ability taxonomies. Four contemporary cognitive taxonomies are examined more closely in terms of psychometric characteristics, such as reliability and validity. At the conclusion of this section, a taxonomy of cognitive abilities is presented. This taxonomy is used in a subsequent section documenting validity evidence for cognitive ability constructs.

Next, in Section III, we examine the research related to conserving human talent in the work force using cognitive ability tests. In this section, we trace events, beginning with the Great Depression, that lead to the development of multi-aptitude test batteries. In particular, we focus on work sponsored by the U.S. Army and U.S. Army Air Forces to develop a variety of batteries for selection and classification purposes during World War II. This section concludes with a description of the current military selection and classification battery.

The next section, Section IV, also focuses on conserving human talent in the work force. In this section, we examine issues and the evidence related to using intelligence and cognitive ability tests to make selection decisions for different subgroups. Also included in this section is a description of Federal regulations enacted to prevent discrimination in job selection. This section concludes with a discussion of the social and legal implications of cognitive ability measurement.

In Section V, we summarize validity data for each cognitive ability construct, using the cognitive ability taxonomy described in Section II. This section includes a description of the types of studies reviewed for this validity summary section. The section concludes with an overall summary of the validity evidence for the nine cognitive ability constructs.

Section VI includes an overall summary of the literature review and presents implications for identifying cognitive ability constructs that may be used to enhance the accuracy of the present Army screening and classification system. This section concludes with a list of cognitive ability constructs that should be considered for supplementing the current military screening battery.

SECTION II

DISCOVERING, MEASURING, AND UNDERSTANDING COGNITIVE ABILITIES

The objective of this section is to develop a cognitive ability construct system that incorporates findings from well over a century of research. It begins with a historical overview of initial attempts to measure intelligence, and descriptions of theories of intelligence that have evolved over the years. Next, we examine two cognitive ability construct systems developed through extensive research and assess their implications for developing a cognitive ability taxonomy. In turn, we examine four widely used multi-aptitude batteries which predict success in occupational or educational settings. The section concludes with a list and definitions of the cognitive ability constructs included in our taxonomy.

Before describing contemporary theories of intelligence, we review very early occupational assessment systems in two societies, China and Greece.

EARLY HISTORY OF OCCUPATIONAL ASSESSMENT

Although psychological testing or, more specifically, the measurement of cognitive abilities appears to be a recent historical phenomenon, DuBois (1970) indicates that modern testing has its roots in very early developments, such as the Chinese civil service examinations. For more than 2,000 years, the Chinese government used an elaborate system of competitive examinations to select personnel for government positions (Bowman, 1989). For example, during the Han Dynasty (206 B.C. to 220 A.D.), written tests were used to assess competency in civil law, military affairs, agriculture, revenue, and geography (DuBois, 1966).

The Chinese selection program during the Ming Dynasty (1368 to 1644 A.D.) evolved into an objective, multi-stage selection system conducted on a nationwide basis. At the local or district level, men vying for public office were given exams which required one day and one night of writing poems and composing essays on two assigned themes. Examinees were also evaluated on penmanship skills. Approximately one to 7 percent passed these tests and went on to the next level. Every three years those passing at the district level were tested in the provincial capital. Testing required nine days and nights and involved writing compositions in prose and verse to reveal the extent of knowledge of the classics. All compositions were transcribed and then evaluated by two independent raters. Approximately 1 to 10 percent of these examinees were considered "promoted scholars" and went on for final testing in Beijing. At the capital level, approximately three percent passed and became eligible for public office.

From this description, it is clear that for thousands of years selection of Chinese public officials was dominated by "ability" testing. Initially the characteristics considered relevant or important for establishing fitness for duty included knowledge of law and current affairs. As the selection system evolved, it appears more emphasis was placed on cognitive abilities such as writing compositions and remembering and interpreting the classics. Today we might refer to these abilities as verbal fluency, reading comprehension, memory, and general reasoning.

If one considers the Chinese to have had a test-dominated society (DuBois, 1970), then the Greeks may be considered to have been a test-influenced society (Doyle, 1974). Although the Greeks lacked a systematic nationwide selection program for public officials, testing to determine vocational fitness was emphasized. For example, in The Republic, Plato notes that "our several natures are not like but different. One man is naturally fitted for one task and another for another" (p. 210). Thus, with individual differences taken into account, the result would be that "more things are produced and better and more easily" (p. 210).

To the Greeks and especially to the Spartans, individual differences were most clearly apparent in the physical abilities area. Young men were constantly evaluated in their skills at the long jump, wrestling, running, and discus throwing. Competition in these areas was used to select and prepare men for the military. In the area of cognitive abilities, philosophers and mentors of the period (e.g., Plutarch, Plato, and Xenocretes) used arithmetic, music, geometry, and astronomy "tests" to screen incoming students. In addition, Plato formulated a theory about the types of intelligence tests that should be used to identify individuals suitable for state office. These facets of intelligence are described in some detail in The Republic. According to our interpretation of Plato's statements, the abilities he recommended assessing include integrative processes, reasoning through complex issues, and memory.

The Greeks also recognized the need for reliable and valid measures of physical and mental abilities. For example, Plato placed special importance on agreement among judges evaluating individual performance; in his opinion, the most effective judges possessed knowledge, good will, and frankness. In addition, tests or measures of physical and mental capacity were standardized to permit more accurate assessment. The validity of these tests appeared to rest "exclusively with estimations of the appropriateness of the content of the test" (Doyle, p. 203), or was established by a procedure that today we refer to as content validity.

From these two examples it is clear that the notion of ability testing has been in existence for thousands of years. In particular, this review of the ancient Chinese and Greek testing systems offers some insight into the types of abilities, aptitudes, and knowledge deemed important for success in their respective societies. Further, Plato, in his discussion of individual differences, offers a rationale for assessing abilities--that is, to match persons with occupations and improve productivity. It was not until recently, however, that issues related to individual differences, intelligence measurement, and the components or facets that comprise intelligence were systematically studied and documented.

HISTORY IN THE MODERN ERA

Bessel

The history of modern research in individual differences begins with an example not from the field of psychology, but instead from the field of astronomy. In 1796 at the Greenwich Observatory, an astronomer's assistant, Kinnebrook, was dismissed by his supervisor, the Astronomer Royal, Maskelyne. The reason for the dismissal was a notable discrepancy in the times of various star movements recorded by Kinnebrook and Maskelyne. Maskelyne naturally assumed that he was correct in his record-keeping and that Kinnebrook was merely being careless. Hence, the hapless Kinnebrook lost his job.

Twenty years later, the incident was researched by another astronomer, Bessel, who had the idea that such differences in records of observations reported by different persons may not be due to error. Instead, he believed that these differences could be a function of what we today call stable individual differences; Bessel labeled them "personal equations." After gathering data on records kept by astronomers, he noted that systematic differences existed between these records, thereby supporting his theory (Dunnette, 1976).

Galton

Researchers in the area of behavioral sciences did not become involved in discovering and assessing individual differences until the late 1800s. The major impetus was the publication of Sir Francis Galton's book, Hereditary Genius (Galton, 1869). In this book, Galton reported his findings on the study of 977 eminent men, who numbered only one per 4,000 people. They came from all walks of life, and included scientists, artists, judges, and writers Galton determined that these eminent men could be classified according to a system of 14 steps or grades. He applied his rating system to their male relatives, starting with father, brother, and son, and continuing on to more remote relatives.

Galton found that as relationship to the proband (eminent individual) became more distant, ratings of eminence declined. Because distant relatives share fewer genes than close relatives, and they show less similarity on the dimension of genius or eminence, Galton concluded that genius was genetically determined. It should be apparent, however, that a major flaw in this study involves the confounding influence of shared environments. That is, close relatives not only share a greater proportion of genes than distant relatives, but are also more likely to share the same or similar environments.

Galton also hypothesized about how to best measure intelligence. For example, noting that all information is received by the senses, Galton reasoned that differences in intelligence could be detected by the measurement of sensory and motor functions. Because mentally retarded persons usually show deficiencies in those processes, this appeared to lend support to Galton's theory. To investigate this hypothesis about the nature of mental abilities, Wilhelm Wundt initiated studies at Leipzig, which became the first laboratory for experimental psychology. Due to the strong influence of the school of structuralism at Leipzig, researchers there emphasized the study of the simplest and most elementary units into which sensory and response functions could be isolated. To do this, they designed laboratory tasks such as von Helmholtz's reaction time paradigm and Fechner's and Weber's psychophysical measures of visual, tactile, and auditory sensitivity. Later, assessment of mental abilities by measuring their components became popular in America. For example, James McKeen Cattell examined physical measures such as grip strength, rate of hand movement, and rate measures including speed of response, rate of perception, and rate of movement (Cattell & Ferrand, 1896).

Ebbinghaus and Wissler

Although Galton's theory of the relationship between the acquisition of knowledge and the process whereby we gain access to this knowledge (i.e., sensory modalities) had intuitive appeal, it proved to be empirically unsupported. Specifically, reports published by the well-known investigator of memory processes, Ebbinghaus (1897), provided evidence refuting the belief that any demonstrable relationship existed between scores on these sensory/psychomotor tests and real-world criteria such as school performance (Dunnette, 1976). He came to his conclusions as a result of testing children in school.

It is of note that Wissler (1901) arrived at the same conclusion after performing a correlational study. Based on his research, Wissler concluded that physical tests show a general tendency to correlate among themselves, but correlate only slightly with tests of mental ability.

TWENTIETH CENTURY BREAKTHROUGHS

Binet and Simon

At about this same time, Alfred Binet also became a major opponent of the school of psychomotor/sensory testing. In their 1895 paper, he and his colleague, Henri, provided an alternative method for measuring intelligence (Binet & Henri, 1895). They believed that good judgment is the defining characteristic of intelligence, and to measure it, they developed tests involving higher, more complex mental functions such as comprehension, reasoning, memory, attention, and adaptation. Binet and Simon published the first intelligence test in 1905 under the auspices of the Parisian government, which commissioned them to develop a method for identifying children who would have difficulties learning in school. Before this time, all classification of mental retardation had been conducted on a purely subjective basis.

Binet and Simon (1905) developed tests consisting of verbal and practical problems, requiring abilities such as judgment, reasoning, and comprehension. Items on the test were designed to be heterogeneous, and thus were more complex than earlier ability tests of perceptual acuity, reaction time, and the like. To Binet and Simon, judgment was the primary characteristic of intelligence, and many of the items on their test were designed to measure this ability. Examples of items of this type include comparing lengths, distinguishing between objects, and completing sentences (Willerman, 1979).

The Binet Test, as it was known, won acclaim and became accepted worldwide, primarily for two reasons: (a) concurrent validity (positive correlations between test scores and rankings of abilities by teachers were found), and (b) predictive validity (the test score predicts the progress of school children, especially for those with low intelligence) (Matarazzo, 1972).

In 1916, Lewis Terman, working at Stanford University in California, translated and revised the Binet-Simon scale into English. The new version was called the Stanford-Binet, and has since gone through two major revisions (Willerman, 1979). The Stanford-Binet tests were designed in such a way that item difficulty levels increased with the subject's age. That is, for each year level there were approximately six test items, resulting in each item having a value of two months of mental age. The items were designed so that 75 percent of the population at the particular age level was able to correctly answer the item. Thus, subjects received two months of mental age credit for each item answered correctly. When mental age is divided by chronological age, the intelligence quotient (I.Q.) results. Thus, children of varying ages can be compared on a relative scale.

The Stanford-Binet was an improvement over the original test because it was applicable to the entire range of human intelligence (from three years of age to adult) and it included more abstract items for the upper levels. Also, it consisted of 90 items as opposed to Binet's original 30. Because it was originally administered orally and required certain activities to be performed in response to some of the items (e.g., unwrapping a piece of candy, comparing weights), test administration time varied greatly. Even the Revised Stanford-Binet, consisting of 129 items (more than half of which are objectively scored), varies in test administration time from 30 to 90 minutes.

The Stanford-Binet was the most widely used individual intelligence test until it was revised in 1937 (and 1960) to become the Terman-Merrill tests (Vernon, 1979). Although the 1937 and 1960 revisions are properly referred to as the Terman-Merrill tests, they are sometimes still called the Stanford-Binet or Revised Stanford-Binet scales. They have been taken by thousands of people over the years, mostly for use in either clinical (diagnosis of learning disabilities/retardation) or educational settings, and occasionally for employment purposes.

Goddard, a key figure in mental testing in the United States, greatly aided the widespread use of the Binet scales, or Stanford-Binet scales, in America. In Goddard's (1912) historical paper describing the Kallikak family, he used intelligence tests to demonstrate the heritability of mental abilities. Based on his success in using the tests, he became a chief proponent of their use throughout society. Tuddenham (1963) gives an account of Goddard's work, noting that it led to the adoption of mental testing in schools, colleges, and military academies everywhere in America.

World War I: Yerkes and Otis

The Binet type of test had disadvantages. The most notable was its cost, in terms of both time and money, because it required individual administration. When the United States entered World War I in 1917, there was an immediate need for objective group testing of large numbers of incoming Army recruits. Robert Yerkes, who was tasked to develop this test, turned to the research of Arthur S. Otis and others. Otis had completed a doctoral dissertation under Terman, wherein he developed a group test called the Otis Self-Administering Test.

Using the Otis test as the basis, Yerkes developed two new intelligence tests, the Army Alpha for the literate and the Army Beta for the illiterate. The Army Alpha required only about 25 minutes for test administration. and appeared to be a stable, reliable measure of cognitive functioning. Components of the tests included verbal, numerical, and reasoning sections.

The Army Alpha and Beta were used to assess 1.7 million men from September, 1917, to January, 1919 (Matarazzo, 1972), resulting in a great wealth of information about the tests and about the groups and individuals completing the tests. Test scores were used to assign troops to the various military jobs requiring different intelligence levels, and to eliminate the non-trainable. The Army Alpha served as a prototype for the development of later tests, especially those for use in industrial selection and placement. Goslin (1963) has estimated that by the 1960s more than 200 million intelligence or achievement tests had been administered in the United States.

Proliferation and Controversy

During the first few decades of the 20th century, modern theories of intelligence were postulated. A major distinguishing feature of these theories was their differing views of the structure or components of intelligence.

Prior to the emergence of many of these theories, an important development occurred in the area of statistical sciences. The method of factor analysis was developed, primarily to lend mathematical support to the various theories. Charles Spearman has generally been regarded as the father of factor analysis, because the groundwork for this technique was laid with his 1904 paper, "General Intelligence, Objectively Determined and Measured." Throughout the rest of his life, Spearman continued to develop and refine the technique.

Much of the early work in factor analysis involved Spearman's tetrad criterion. This famous theorem states that if certain relationships exist among the correlations of a set of variables, then each variable can be described in terms of a general factor $\underline{g}$ and $\underline{a}$ specific factor $\underline{s}$. More specifically, to describe a set of variables according to Spearman's two-factor theory, all tetrads must vanish, as follows:

$$r_{jk}r_{lm} - r_{lk}r_{jm} = 0 \text{ where } j, k, l, m = 1, 2, \ldots n; \text{ and } j \neq k \neq l \neq m$$

When these conditions hold, the two-factor pattern is assumed to hold true. In this way, Spearman's theory can be statistically verified.

The purpose of factor analysis is to represent observed variables in terms of several underlying hypothetical constructs, or "factors." This smaller number of factors should not only extract a maximum amount of variance from the variables, but also accurately reproduce the observed correlations between them. The cluster of variables making up a factor is said to load on the factor. Hence, factor analysis is an attempt to describe observed data parsimoniously, and usually serves as an exploratory device.

It is important to note that factor analysis is not able to identify basic dimensions in fields such as psychology. That is, the technique is purely a statistical one, and the factors emerging have no psychological meaning in and of themselves. It is the responsibility of the investigator to allocate meanings or labels to the factors, based upon the variables loading on them. When it first became popular, factor analysis was perceived by some as a kind of mystical method that could be implemented to find "true" latent dimensions of behavior. However, as Kelley (1940, p. 120) has pointed out:

There is no search for timeless, spaceless, populationless truth in factor analysis; rather, it represents a simple straightforward problem of description in several dimensions of a definite group functioning in definite manners, and he who assumes to read more remote verities into the factorial outcome is certainly doomed to disappointment.

In the 20 years following the publication of Spearman's first paper on factor analysis, a great deal of research ensued on the technique and its application to psychological theories of intelligence. Spearman continued to contribute to the effort; others who were active include Karl Pearson, L. L. Thurstone, Cyril Burt, Godfrey Thomson, J. C. Maxwell Garnett, and Karl Holzinger. According to Harman (1976), the bulk of the work at this time addressed the existence of _g_, the study of sampling errors of tetrad differences, and computational methods to derive a single general factor.

As will become apparent in later sections of this report, Spearman's two-factor theory of intelligence found statistical support through factor analysis, as he discovered that all variables could be resolved into linear expressions involving only a general factor and specific factors unique to each variable or test. When it was later realized that group factors were also important, the theory changed but the method through which the factors were derived remained essentially the same. Hence, as scientific theory evolved, factor analysis proved to be flexible enough to adapt to changes and still serve as a useful statistical tool in theory testing.

Pearson's (1901) main contribution was in setting forth the method of principal axes, a statistically optimal solution in which each factor is determined in sequence, so that at each successive stage the factor accounts for a maximum amount of variance. A major contribution of Thurstone (1938a) was to popularize the method of multiple factor analysis. He concluded from this type of analysis that Spearman's g factor was only a second-order factor--that is, the result of intercorrelations among first-order factors. Thurstone contended that intelligence was composed of many primary factors. He originally identified 12 primary mental abilities, using his multiple factor analysis technique.

THEORIES OF INTELLIGENCE

Measurement of a construct requires some agreement among experts concerning the nature of the construct. When the measurement of intelligence became increasingly sophisticated at the beginning of the 20th century, various theories of intelligence were proposed. These theories are important to the discussion of measurement, because inherent in each are guidelines regarding appropriate methodologies and criteria. The following paragraphs outline selected major theories of intelligence developed during the early part of this century.

Spearman

In 1904, Charles Spearman developed and popularized the unitary or monarchic theory of mental abilities, emphasizing the general factor that he called g. He conceived of g as an innate general mental energy underlying all cognitive processes. His theory also included s, or specific factors, which were learned, rather than innate, and were associated with the different intelligence tests. Evidence for the two-factor theory was provided in the finding that all the tests showed positive intercorrelations. Tests were assumed to show this intercorrelation to the extent to which they were saturated with g, and this part of the total unit variance was termed the communality. The residual, or unique variance, comprises both specific variance due to the variables and error variance due to unreliability. It is important to note that Spearman's tests differed greatly from those used today in generating and testing this theory. Rather, his tests measured subjects' sensory discrimination power, including hearing, sight, and touch (Spearman, 1904). For example, subjects were asked to determine differences in pitch, hue, and weight of various stimuli.

Anastasi (1983) noted that Spearman believed that s factors do not add to the explainable variance because they operate only in specific tests. Their usefulness, then, comes not from their predictive validity but from using them to obtain more nearly pure measures of g. Since the s factors have been parceled out, one is left with g, the factor underlying all abilities.

Essential to Spearman's theory were the two components that he believed comprise intelligence. The first is the eduction of relations, which is the ability to extract a relationship between two givens; an example is the induction of the relationship "synonym" when given the terms "small" and "little." The second is the eduction of correlates, which refers to the capacity to apply the educed relation to a different situation--for example, to fill in the missing synonym in "large, _____." This latter component is usually known as reasoning by analogy.

Thorndike

In the early part of this century, Spearman's model of the nature of intelligence dominated most psychologists' conceptualizations of mental abilities. Beginning in the 1920s, however, it appeared to some researchers that Spearman's theory was overly simplified. E. L. Thorndike was another opponent of Spearman's *g* theory; his view was that the mind consisted of many bonds or connections, and that intelligence test items sample these bonds. This sampling theory hypothesized that there were individual differences in the ability to form different types of connections, and that these differences were innate and could not be trained or otherwise altered. The different types of connections were postulated to be independent specific abilities, such as verbal ability or spatial ability, each with its own neural substrate.

Thorndike's theory assumed that tests intercorrelate to the extent that they sampled or drew upon common bonds. Some bonds tend to cluster, and these might form group factors such as *verbal*, *spatial*, or *quantitative*. By positing this theory of sampling, the positive correlations found between intelligence tests could be explained without invoking a *g* concept.

The major evidence supporting this theory came from a study conducted by Thorndike, Bregman, and Cobb (1927). The purpose of the experiment was to examine differences between informational or associative thinking tasks (thought to represent acquired knowledge) and reasoning on inferential thinking tasks (thought to represent innate abilities). Three informational and three reasoning tests were administered to a sample of 250 eighth-grade males. The informational tests intercorrelated .60, and the reasoning tests intercorrelated .54. The mean correlation between the two sets of tests was .60. Thorndike believed that this was evidence that the informational tests and the reasoning tests were equivalent measures of intelligence. Since the proposed distinction between innate versus acquired abilities was not supported, he concluded that this outcome supported the theory that intelligence represents only the total number of connections in the mind, regardless of source. In Thorndike's opinion, this evidence fails to substantiate Spearman's theory.

Later investigators in the area of experimental psychology found that the sampling theory did not adequately explain individual differences in mental abilities. The number of S-R associations that would have to exist to uphold the theory appears to be almost infinite, meaning that no human could learn as many S-R associations as needed to function adequately. Another problem is the assumption that these connections are dependent upon the synapses between neurons. Physiologically, this does not appear to be possible (Vernon, 1979).

Thurstone

Some of the strongest criticisms of Spearman's theory of intelligence came from L. L. Thurstone, who developed the technique of multiple factor analysis. Upon administering a large battery of cognitive tests (56) to 240 University of Chicago students, he found 12 factors resulting from a centroid analysis with an orthogonal rotation to a final solution (Thurstone, 1938a). Of course, Spearman's theory of *g* would have predicted the finding of only one general factor. Thurstone's method of analysis differed from Spearman's in two ways. First, the orthogonal rotation method used by Thurstone did not allow factors to be correlated. Second, the multiple factor analysis techniques he used enabled him to identify Spearman's *g* factor as a second-order factor, as explained previously.

The lack of interpretability of some of Thurstone's factors, however, warranted further examination into the nature of the structure of intelligence. A subsequent investigation of 215 high school seniors yielded nine factors. Four of these were combined, resulting in Thurstone's final six factors of mental ability: (1) verbal comprehension, (2) number, (3) word fluency, (4) space, (5) associative memory, and (6) inductive reasoning (Thurstone, 1938b). Another ability, later replicated in several studies, was that of perceptual speed. Hence, Thurstone concluded that intelligence does not comprise a single unitary factor but, rather, several types of abilities. Performance on a given task requires a mixture of these primary mental abilities in some proportion, analogous to the manner in which primary colors can be combined to yield any color of the spectrum.

It should be noted that although Thurstone found very low correlations between subtests and, therefore, concluded that the primary mental abilities were orthogonal, these conclusions were later discovered to be an artifact of the sampling procedure. The homogeneous sample of the University of Chicago students used by Thurstone in his initial research was restricted in intellectual range to the upper t2ail of the distribution, thereby attenuating test score intercorrelations. This may be one reason why Thurstone was able to extract several factors rather than one large common factor. Indeed, in working with a more representative sample of younger students, he found the primary factors to be oblique rather than orthogonal (Thurstone, 1938b).

Burt and Vernon

Hierarchical models of abilities have been proposed by both Burt (1949) and Vernon (1961). These models assume the existence of *g* and *s* factors, as well as intermediate factors known as major and minor group factors. Support for this type of theory comes from factor analytic results in which a small number of correlated group factors have been obtained. Verbal/educational and practical/mechanical are two of the group factors that often emerge when ability test data are factor analyzed (Vernon, 1961). These two major factors are further subdivided into verbal and numerical subfactors, and mechanical-informational, spatial, and psychomotor ability subfactors, respectively. At the next level of the hierarchy, the subfactors can also be divided. This continues until the lowest level, corresponding to Spearman's _s_ factor, is reached (Anastasi, 1983).

Positing the existence of group factors in addition to _g_ represents an attempt to obtain a better fit to the empirical data. Vernon's and Burt's approaches are, thus, both applied and pragmatic. In developing the theory, both were cautious in generalizing from statistical factors to psychological factors, avoiding reification of the factors. It is also notable that Burt used factor analysis to _test_ his theory, rather than to generate it. This demonstrates another way in which Burt's theory differs from Thurstone's. Thurstone started with the use of factor analysis and the finding later led to the development of his theory; Burt began with a theory, and used factor analysis in later stages.

Vernon has, however, identified two problems with the use of hierarchical models. First, he admits that other factor-analytic approaches (e.g., Thurstone's centroid method, Hotelling's principal components' technique) are more mathematically precise because they more accurately depict the relationships among the different measures. Second, the specific factors assumed to be lowest on the hierarchy appear to be of little use because by definition they lack real-life variance. Vernon labels a factor "specific" if it contributes less than 5 percent of the variance of some criterion such as educational or occupational proficiency. He suggests that broader factors are more important for the applied psychologist, especially because of their empirically demonstrated predictive validity. Examples of these specific factors lacking in predictive validity are rote memory (as measured by Thurstone), manual dexterity, coordination, and sensory-motor factors (Vernon, 1964).

Guilford

As Dunnette (1976) has pointed out, initial theories of intelligence assumed the existence of one underlying mental factor, then moved to the assumption of several, and finally to the postulation of many. J. P. Guilford's (1967) Structure-of-Intellect (S-I) theory is an example of the last, a theory positing the existence of 120 or more factors. These factors result from a model using three dimensions of classification:

1. Operations - what the respondent does; Guilford's five hypothesized operations are cognition, memory, convergent thinking, divergent thinking, and evaluation.

2. Contents - the nature of the materials or information on which operations are performed; Guilford's four hypothesized content factors are semantic, figural[1], symbolic, and behavioral.

---

[1]In the latest version of the structure-of-intellect model, the figural category has been replaced by visual and auditory content categories (Guilford, 1982).

3. <u>Products</u> - the form in which information is processed by the respondent; Guilford's six hypothesized product factors are units, classes, relations, systems, transformations, and implications.

For example, in this system, verbal comprehension corresponds to cognitive operations of semantic (content) units (products). One factor results from each combination of these classifications. As of 1971, 98 of these 120 aptitude factors had been identified by Guilford and his associates (Guilford & Hoepfner, 1971).

Guilford (1981) derived his factors through the use of orthogonal rotation, preferring this method for several reasons. First, he noted that Thurstone used orthogonal rotations and found factors that were generally replicated across different samples, different times, and different measures (see French, 1951). Second, he pointed out that orthodox oblique rotational methods failed, in that they resulted in uninterpretable results that could not be replicated. Because Guilford considers his 120 factors to be orthogonal, he rejects the utility of <u>g</u> and of hierarchical relationships among the factors. He admits, however, that the third-level factors appear to be the most general; these are the operation categories. Because they are the next most discriminable, the content categories appear next, followed by the product classes. Guilford warns that this hierarchical model as applied to S-I theory is still flawed; there is a lack of space for third-order content and product factors and their subsidiary second-order abilities.

It is noteworthy that Guilford's theory has been guided from the beginning by an a priori theoretical model. He developed this structure-of-intellect theory first, then concentrated on the development of tests to measure the specific components hypothesized in his theory. It is theory, then, that has guided all of Guilford's test development efforts. Binet, on the other hand, was a strong opponent of this approach; his test was developed in the opposite manner. Binet began with empirical data resulting from the administration of his test, and from these data he developed his theory of intelligence. Hence, the distinction between the approaches of Guilford and Binet can be perceived as a distinction between deductive and inductive approaches, respectively.

<u>Eysenck</u>

Eysenck's (1953) model of intellect is similar to Guilford's. It consists of three dimensions, two of which appear to overlap with the structure-of-intellect model. Eysenck's "mental processes" include reasoning, memory, and perception; they are similar to Guilford's operations. Eysenck's "test materials" include verbal, numerical, and spatial; these are similar to what Guilford calls contents. Eysenck's and Guilford's models differ, however, with respect to the third dimension. The idea of products seems unimportant to Eysenck; so, instead he substitutes "quality," which incorporates the concepts of mental speed and power. This emphasizes the notion that speed and power are fundamental to all mental work but are qualified by both mental processes and test materials. Eysenck (1967) points out that his model retains the concept of g in a hierarchical structure, and that the major source of variation is mental speed, averaged over all processes and materials.

Summary

At this point, it will be useful to review the recent theories of the nature of intelligence, and their implications for modern views of cognitive abilities. Anastasi (1983) has created a taxonomy or framework into which most of the previously discussed theories fit. There are four major models in her classification scheme: two-factor, multiple-factor, facet, and hierarchical. The latter three models postulate the existence of a number of factors, but differ in terms of the relationships specified between these factors.

The chief representative of the two-factor theory is, of course, Spearman's theory of g and s. In brief, g is the factor measured by all ability tests to some degree, and s denotes the factor specific to the individual test.

When Thurstone attempted to replicate Spearman's two-factor structure, he found that a broader group of factors better explained the correlations between tests. Hence, Thurstone became a proponent of the multiple-factor model of intelligence. It is important to remember that the tests used by the two men differed substantially. Spearman used tests of discriminative ability (measuring subjects' sensitivity with respect to auditory, visual, and tactile stimuli), while Thurstone used written tests much more similar to the kind that his subjects may have been exposed to (e.g., simple arithmetic tests, sentence completion, paragraph comprehension).

Although not mentioned in the Anastasi (1983) paper, Thorndike would probably also fit into the category which Anastasi has labeled multiple-factor. His sampling theory allowed for the existence of clusters of bonds forming group factors that could be considered similar to the primary mental abilities found by Thurstone.

Vernon's theory is the best-known hierarchical model of cognitive abilities. Anastasi notes that this model permits the integration of Spearman's g with the abilities found in the multiple-factor models. The hierarchical models begin with g at the broadest level and decompose that factor until reaching the lowest level, specific factors. The primary mental abilities probably fall at about the third level of a hierarchical model.

The structure-of-intellect model proposed by Guilford is an excellent example of a facet theory. Guttman (1958) was the first to apply this label to this type of trait model, and described it as a design in which each test could be defined by specifying the facets or dimensions that applied to it. Because Eysenck's theory can be considered a modification of Guilford's, it is also a facet theory.

Modern views of the nature of intelligence have been influenced by all of these theories. Of course, there is still no agreement among "experts" in the field of cognitive-ability testing as to which factors constitute the construct called intelligence. There seems to be some consensus that intelligence comprises a number of specific abilities. Although there is disagreement concerning the existence of a g factor, almost all researchers acknowledge the existence of separate, relatively independent specific factors.

SPECIFICATION OF COGNITIVE ABILITIES

As the preceding discussion indicates, trends in intelligence theory and measurement have shifted from Spearman's conceptualization of a single, unitary trait, to one involving several facets of intelligence. Whether these facets or cognitive abilities are independent from one another or can be systematically ordered in some hierarchical fashion will not be debated here. The more relevant issue concerns the numbers and types of intellectual facets or cognitive abilities that may be reliably measured and used to predict work performance outcomes. To address this issue, we backtrack somewhat and focus on Thurstone's research results and the subsequent research conducted by Guilford and researchers from Educational Testing Services.

Guilford Revisited

Model Development and Empirical Results. As noted previously, results from Thurstone's (1938a) large factor-analysis study suggested that six primary mental abilities could be differentiated. Subsequent research by Thurstone and his students (e.g., Bechtoldt, 1947; Taylor, 1947; Thurstone, 1944) provided verification of the primary mental abilities and suggested the possibility of more ability factors. Guilford and others, while conducting pilot selection research for the U.S. Army Air Force Aviation Psychology Research program, determined that more specific cognitive ability factors than those identified by Thurstone could be demonstrated and differentiated. Many of Thurstone's primary ability measures, among others, were adapted for use in pilot selection. Factor analysis results of test intercorrelations indicated that perceptual, space, memory, and reasoning abilities could be further differentiated into more specific subcomponents. For example, results from this study raise questions about the distinction between visualization and spatial ability. And within the visualization factor, distinctions between two-dimensional and three-dimensional rotation were also examined (Guilford & Lacey, 1947).

Following World War II, Guilford continued investigating the factor structure of human intellect as director of the Aptitudes Research Project at the University of Southern California. By 1955, research in the area led to the conclusion that about 40 specific cognitive-ability factors could be identified.

To establish a pattern of the relationships among the growing list of demonstrated ability factors, Guilford proposed the Structure-of-Intellect

(S-I) theory and model (Guilford & Hoepfner, 1971). As noted previously, the S-I theory posits the "existence" of some 120 distinct cognitive abilities, although Guilford notes that even more abilities may be identified using different stimulus presentation modalities (e.g., verbal vs. audio). From approximately 20 years of research in the structure of intellectual abilities, Guilford and his colleagues have established support for the S-I theory and model.

Briefly, the procedures followed to demonstrate the existence of distinct ability factors included developing measures of the six operation-by-content product factors. For example, these six measures may include cognition of figural (a) units, (b) classes, (c) relations, (d) systems, (e) transformations, and (f) implications. According to Guilford and Hoepfner, "this strategy has been a fairly good one, since it has been found generally more difficult to differentiate abilities differing only as to products" (p. 10).

Constructed measures of the six operation content-by-product factors were administered to samples of military recruits, military officer candidates, high school students, or elementary students. Test scores were intercorrelated and factor analyzed, using a targeted principal factor technique accompanied by rotation to an orthogonal solution. Using these procedures, results documented in 41 technical reports indicate that 98 of the proposed 120 distinct ability factors have been demonstrated. Guilford and Hoepfner (1971) noted that because of the factor analytic approach used, the independence of ability factors differing in products has been demonstrated whereas there is much less information about the independence of ability factors differing in content or operations. Cumulative results from the 41 technical studies are reported in Table 1, indicating which ability factors have been demonstrated and which require further research.

From this table, it can be seen that most of the ability factors not yet demonstrated fall within the behavioral content categories. Of the 22 factors lacking empirical support, 18 are behavioral content factors. According to Guilford, these factors correspond to Thorndike's notion of social intelligence. This category of ability factors includes non-figural and non-verbal information involved in human interactions which encompass "attitudes, needs, desires, moods, intentions, perceptions and thoughts of others and of ourselves" (Guilford & Hoepfner, 1971, p. 21). Tests designed to measure these abilities generally involve drawings or photographs of facial expressions. Respondents are asked to select the correct captions, identify the correct caption for a photograph, or indicate a series of facial expressions that go together in some way or tell a story.

As results from both the table and Guilford himself indicate, developing measures that assess abilities to comprehend, evaluate, or understand human interactions and cues is difficult at best. Although Guilford and his colleagues continue to investigate the possible existence of a "social intelligence factor" (O'Sullivan & Guilford, 1976; O'Sullivan, 1983), evidence from other sources suggests that it is difficult to reliably measure or find support for such a factor (Woodrow, 1939; Ekstrom, French, & Harman, 1979; Frederiksen, Carlson, & Ward, 1984). Recall that, earlier in this report, we reported that these types of abilities (e.g., social competence) would be excluded from our operational definition of the cognitive ability domain.

Table 1

Structure-of-Intellect Factors[a] That Have Been Demonstrated (Uppercase Trigrams) and Those That Have Not Been Demonstrated (Lowercase Trigrams)

Operation Categories	Content Categories: Figural	Symbolic	Semantic	Behavioral	Number Known
Cognition	CFU CFC CFR CFS CFT CFI	CSU CSC CSR CSS CST CSI	CMU CMC CMR CMS CMT CMI	CBU CBC CBR CBS CBT CBI	24
Memory	MFU MFC MFR MFS MFT MFI	MSU MSC MSR MSS MST MSI	MMU MMC MMR MMS MMT MMI	mbu mbc mbr mbs mbt mbi	18
Divergent Thinking	DFU DFC dfr DFS DFT DFI	DSU DSC DSR DSS DST DSI	DMU DMC DMR DMS DMT DMI	DBU DBC DBR DBS DBT DBI	23
Convergent Thinking	nfu NFC NFR nfs NFT NFI	nsu NSC NSR NSS NST NSI	NMU NMC NMR NMS NMT NMI	nbu nbc nbr nbs nbt nbi	15
Evaluation	EFU EFC EFR EFS EFT EFI	ESU ESC ESR ESS EST ESI	EMU EMC EMR EMS EMT EMI	ebu ebc ebr ebs ebt ebi	18
Number Known	27	29	30	12	98

Note: From The analysis of intelligence by J. P. Guilford and R. Hoepfner (1971), p.55. New York: McGraw-Hill. (Copyright 1971 by McGraw-Hill.) Reprinted by permission.

[a]Factor Codes: Factors are designated by letters for each parameter which appear in the following order: Operation, Content, Product.

Product Codes are as follows: U = Unit, C = Class, R = Relation, S = System, T = Transformation, I = Implication.

Unique Features of the S-I Model. Perhaps one of the most distinguishing features of the S-I theory and model involves the Divergent Thinking factors. According to Guilford, problems related to leadership in the Army Air Force, emerging during the Aviation Psychology Research Program, led to the generation of hypotheses about divergent thinking factors (Guilford & Hoepfner, 1971). Prior to the formulation of these ability factors, measures of intelligence such as the Stanford-Binet emphasized convergent thinking, or finding a single correct answer to a problem. As Dunnette (1976) pointed out, "It is no surprise that Binet missed an additional important aspect of human ability, divergent thinking. He based his selection of items on ratings of non-test behaviors that failed to emphasize divergent thinking abilities" (e.g., school performance, teachers' ratings) (p. 480).

The divergent thinking factors are designed to assess or uncover creative thinking processes, such as originality, flexibility, and fluency. So far, results from research involving measures of these abilities indicate that they correlate with similar types of measures but seldom can be used to predict the ability to develop innovative or creative products in a productive endeavor. As Dunnette noted, however, two aptitudes have been identified that are independent of traditional intelligence test scores (at least among college students) and that are predictive of behaviors related to creative production. They are Ideational Fluency and Preference for Complexity-Asymmetry over Simplicity-Symmetry.

Ideational Fluency involves the capacity for generating ideas about a particular topic, theme, or picture. For such measures, subjects' responses are typically scored on quantity of output and not quality. Carefully conducted investigations using Ideational Fluency measures indicate that test scores may be used to predict achievement and accomplishment in such areas as performing arts, literature, mathematics, science, crafts, social science, and leadership in school and in college (Csikszentmihalyi & Getzels, 1970; Hocevar, 1980; Singer & Whiton, 1971; Wallach & Wing, 1969).

The second type of measure, Preference for Complexity-Asymmetry, has best been measured by Barron and Walsh with a test of heterogeneous line drawings selected empirically to differentiate between artists and non-artists. Measures of this construct have proven to be predictive of creative behaviors in other fields such as writing, architecture, and scientific research (Dellas & Gaier, 1970). Although it is still true that little is known about the most effective means of measuring this ability factor, it is clear that a few of these measures are tapping something other than traditional intelligence (Dellas & Gaier, 1970; Getzels & Jackson, 1962; Hocevar, 1980; Wallach & Wing, 1969).

Evaluation of the Model. Using Guilford's S-I theory and model to establish a cognitive construct taxonomy, however, presents some problems. The model has been criticized as both logically and methodologically problematic. On logical grounds, Carroll (1972) stated that because Guilford's factors are claimed to be orthogonal or independent from one another, the postulated classification structure imposed is really not required. On metrodological grounds, Horn (1967) indicated that the factor analysis procedures used by Guilford and his colleagues have been too subjective or have permitted easy confirmation of hypothesized factors. Horn and Knapp (1973) also argued that Guilford's factor-analytic procedures are too subjective because they permit confirmation of any hypothesized set of factors, even those derived randomly.

Carroll summarized additional evidence (Harris & Harris, 1971; Haynes, 1970) demonstrating the lack of support for the structure-of-intellect model. In the Haynes study, two of the best tests representing each of 17 of the most clearly established factors were administered to college students (N = 200) and the resulting test scores were factor analyzed; all but six tests demonstrated factor loadings of .30 or higher on a general factor. In addition, although 12 group factors could be identified, the distinction between specific types of products, contents, and operations, was blurred. Further, Harris and Harris' reanalysis of a subset of Guilford's factors, again using a more objective factor-analysis technique, yielded more traditional ability factors, such as verbal comprehension, arithmetic facility, deductive reasoning, inductive reasoning, and word fluency.

A final caveat for using the structure-of-intellect model as a guide to establish a cognitive ability taxonomy concerns the linkage between the S-I factors and dimensions or categories of work performance. According to Dunnette (1976):

> The structure-of-intellect model has been internally oriented, making little or no contact with the real world of human work performance; as such, the theory and the tests designed to test the theory are of little direct use for further elaborating and understanding of the patterns of human attributes important for an understanding of work performance in organizational settings (p. 480).

Factor-Referenced Cognitive Tests (Educational Testing Service)

Development of the Kit. Another attempt to lend structure to the cognitive abilities domain is provided by researchers at the Educational Testing Service (ETS) (Ekstrom, French, & Harman, 1979). Work originated by French (1951) was also based on Thurstone's primary mental ability factors. French compiled a summary of factor-analytic studies which yielded a list of 59 different ability factors that had been sufficiently identified to receive names (Carroll, 1982). Results from this analysis led to the development of the Kit of Selected Tests for Reference Aptitude and Achievement Factors (French, 1954). This Kit contained marker tests or measures for what French considered 16 well-established cognitive abilities and achievement factors. Further research conducted in the 1950s, designed to identify additional cognitive ability factors, led to the development of a revised battery, the Kit of Reference Tests for Cognitive Factors (French, Ekstrom, & Price, 1963). More recently, the Kit was revised, yielding a battery of marker tests for 23 well-established cognitive ability factors (Ekstrom, French, Harman, & Durman, 1976).

The theory and procedures underlying the identification of the cognitive ability factors measured by the Kit differ from Guilford's approach. For example, in the 1976 Kit test manual, the authors state specifically that "cognitive factors resist classification by any rigid taxonomy such as Guilford's Structure of Intellect model" (Ekstrom et al., 1976, p. 3).

A second difference in identifying the cognitive abilities for inclusion in the Kit involves the data used to support or demonstrate the usefulness of a particular factor. Unlike Guilford, who in general utilized only data he or his colleagues collected, these authors required data from independent sources to demonstrate the existence of a cognitive ability factor. In fact, the current Kit includes cognitive ability factors suggested in research by a number of investigators, such as Guilford, Royce, Cattell, and Carroll. According to Ekstrom and her colleagues (1979), a factor was included and marker tests developed if the factor had been identified in at least two different laboratories. Thus, "no one researcher's factors are considered established unless they have been replicated by others" (p. 8).

Specific Abilities Assessed in the Kit. Because this set of factors represents, perhaps, the most comprehensive list of established cognitive abilities, a closer examination is appropriate. Below we provide a list and brief description of each factor included in the most recent Kit of-Factor-Referenced Cognitive Tests (Ekstrom et al., 1976). Also included are the sources used to identify each factor. Note that several of Guilford's factors are represented, as are all of Thurstone's primary mental abilities.

1. Flexibility of Closure - the ability to hold a given percept or configuration in mind so as to disembed it from other well-defined perceptual material.

   Source: Guilford's convergent production of figural transformations (NFT) and Thurstone's Closure 2 - flexibility of closure.

2. Speed of Closure - ability to "take in" a perceptual field as a whole, to "fill in" unseen portions with likely material, and thus to coalesce somewhat disparate parts into a visual percept.

   Source: Guilford's cognition of figural units (CFU) and Thurstone's Closure 1 - speed of closure.

3. Verbal Closure - the ability to solve problems requiring the identification of visually presented words when some of the letters are missing, scrambled, or embedded among other letters.

   Source: Guilford's cognition of symbolic units (CSU).

4. Associational Fluency - the ability to rapidly produce words which share a given area of meaning or some other semantic property.

   Source: Guilford's divergent thinking of semantic relations (DMR).

5. Expressional Fluency - the ability to rapidly think of word groups or phrases. Expressional fluency differs from ideational fluency in requiring rephrasing of ideas already given instead of the production of new ideas. Based on recent research, there appears to be little support for this factor (Ekstrom et al., 1976).

   Source: Guilford's divergent thinking of semantic systems (DMS).

6. Figural Fluency - the ability to quickly draw a number of examples, elaborations, or restructurings based on a given visual or descriptive stimulus. This may be a figural form of ideational fluency.

   Source: Guilford's divergent thinking of figural units, implications, and systems (DFU, DFI, and DFS).

7. Ideational Fluency - the facility to write a number of ideas about a given topic or examples of a given class of objects; ability which provides for rapid production of ideas fitting a given specification.

   Source: Guilford's divergent thinking of semantic units (DMU).

8. Word Fluency - the facility to produce words that fit one or more structural, phonetic, or orthographic restrictions that are not relevant to the meaning of words; this factor accounts for the ability to rapidly produce words fulfilling specific symbolic or structural requirements.

   Source: Guilford's divergent thinking of symbolic units (DSU) and Thurstone's W - word fluency.

9. Induction - the kinds of reasoning abilities involved in forming and trying out hypotheses that will fit a set of data.

   Source: Guilford's cognition of symbolic classes and systems (CSC and CSS) and of figural classes (CFC).

10. Integrative Processes - the ability to keep in mind simultaneously or to combine several conditions, premises, or rules in order to produce a correct response.

    Source: Guilford's memory of symbolic relations (MSR).

11. Associative Memory - the ability to recall one part of a previously learned but otherwise unrelated pair of items when the other part of a pair is presented.

    Source: Guilford's memory of symbolic implications (MSI) and Thurstone's M - rote memory.

12. Memory Span - the ability to recall a number of distinct elements for immediate reproduction.

    Source: Guilford's memory of symbolic units (MSU).

13. Visual Memory - the ability to remember the configuration, location, and orientation of figural material.

    Source: Guilford's memory of figural units, classes, and relations (MFU, MFC, and MFR).

14. Number Facility - the ability to perform basic arithmetic operations with speed and accuracy. This factor is <u>not</u> a major component in mathematical reasoning or higher mathematical skills.

    Source: A subfactor of Guilford's memory of symbolic implications (MSI) and Thurstone's N - numerical facility.

15. Perceptual Speed - speed in comparing figures or symbols, scanning to find figures or symbols, or carrying out other very simple tasks involving visual subfactors such as form discrimination and symbol discrimination. These are, however, more usefully treated as a single concept for research purposes.

    Source: Guilford's evaluation of figural and symbolic units (EFU and ESU), and Thurstone's P - perceptual speed.

16. General Reasoning - the ability to select and organize relevant information for the solution of a problem, including that of a mathematical nature.

    Source: Guilford's cognition of semantic systems (CMS).

17. Logical Reasoning (Deduction or Syllogistic Reasoning) - the ability to reason from premise to conclusion or to evaluate the correctness of a conclusion.

    Source: Guilford's evaluation of semantic relations (EMR) and Thurstone's D - deduction.

18. Spatial Orientation - the ability to perceive spatial patterns or to maintain orientation with respect to objects in space.

    Source: Guilford's cognition of figural systems (CFS) and Thurstone's S - space.

19. Spatial Scanning - speed in visually exploring a wide or complicated spatial field.

    Source: Guilford's cognition of figural implications (CFI).

20. Verbal Comprehension - ability to understand the English language.

    Source: Guilford's cognition of semantic units (CMU) and Thurstone's V - verbal comprehension.

21. Visualization - the ability to manipulate or transform the image of spatial patterns into other arrangements; ability to manipulate visual precepts and thus to "see" how things would look under altered conditions.

    Source: Guilford's cognition of figural transformations (CFT).

22. Figural Flexibility - the ability to change set in order to generate new and different solutions to figural problems.

    Source: Guilford's divergent thinking of figural transformations (DFT).

23. Flexibility of Use - the mental set necessary to think of different uses for objects.

    Source: Guilford's divergent thinking of figural transformations (DFT) and convergent thinking of semantic transformations (NMT).

Although the 1976 version of the Kit (Ekstrom et al., 1976) is very similar to the earlier Kit of Reference Tests for Cognitive Factors (French et al., 1963), some differences do exist. First, four factors--Sensitivity to Problems, Length Estimation, Mechanical Knowledge, and Originality--were deleted, either because they were too narrow or because subsequent data failed to replicate them. Second, the Flexibility of Use Factor (23) was created by combining two factors appearing in the earlier battery, Spontaneous Semantic Flexibility and Semantic Redefinition. And third, recent efforts to establish or demonstrate new factors resulted in the inclusion of Verbal Closure (3), Figural Fluency (6), Integrative Processes (10), and Visual Memory (13). Another proposed factor, Concept Formation/Attainment, proved to be inadequately demonstrated and was, therefore, excluded from the battery (Ekstrom et al., 1976).

In their 1979 monograph, Ekstrom and her colleagues provided a summary of the evidence to date as to the independence of the 23 cognitive ability factors. According to these data, Flexibility of Closure (1) and Speed of Closure (2) are not easily distinguishable although Verbal Closure (3) appears distinct from the two. The five fluency factors (4, 5, 6, 7, and 8) appear very closely related. General Reasoning (16) and Integrative Processes (10) are difficult to separate from other reasoning factors. Spatial Orientation (18) and Visualization (21) are not easily distinguishable. These findings suggest that the following list of factors may be used to represent the cognitive abilities area:

1. Flexibility and Speed of Closure
2. Fluency
3. Induction
4. Associative (Rote) Memory[3]
5. Span Memory[2]
6. Number Facility
7. Perceptual Speed
8. Deduction (Logical Reasoning)
9. Spatial Orientation and Visualization
10. Spatial Scanning
11. Verbal Comprehension

As Dunnette (1976) notes, it is remarkable that the years of factor-analytic research have added only a few constructs to Thurstone's list of seven primary mental abilities (the original six primary ability factors plus a perceptual speed factor identified in subsequent research efforts).

<u>Evaluation of the Two-Construct Systems</u>. The purpose of reviewing Guilford's theory and model and the battery constructed by Ekstrom and her colleagues is to help formulate a cognitive ability taxonomy for use in the present study. It may be useful to consider an issue related to the procedures used by both research teams to construct or establish their respective taxonomies. Both teams relied heavily upon factor analysis to isolate specific, independent cognitive factors. Problems related to Guilford's factor-analytic procedures have already been noted. A broader issue relates to the assumption underlying resulting factors--that is, obtained factors are assumed to be invariant or replicable when applied to similar or slightly different populations, using the same tests. Evidence reported by the Ekstrom group in 1979 indicated that this is not always the case.

Over the years researchers have generated hypotheses about variables that may alter the obtained cognitive ability factor structure. Anastasi (1983) argued that the differentiation hypothesis may explain some of the differences in factor structure. According to this hypothesis, in early childhood intelligence is relatively undifferentiated whereas it becomes more specialized into distinct group factors as one moves from childhood to adolescence. In addition, different factors will emerge at the high school level if the population includes students with an emphasis on academic versus technical course work. Further, the ability level of the population affects the number of factors emerging. For groups of higher ability level, or homogeneous samples with respect to intelligence levels, there is greater differentiation among ability factors. Moreover, the type of resulting factors varies depending upon the area in which the sample excels. For example, for a high verbal group, several clearly identifiable verbal

---

[2]The monograph, in describing the results for the Verbal Memory factor, does not provide information about how this factor relates to the other memory factors. Until such information becomes available, Verbal Memory will not be treated as a separate factor.

factors will emerge; for a group that excels in spatial ability, two or more spatial factors may emerge while only a single verbal factor appears.

The upshot is that the number and type of factors emerging from the data may vary depending upon characteristics of the sample, such as education level, ability level, and educational course focus. Research designed to identify general cognitive constructs may lead to different conclusions about the structure of cognitive abilities if samples representing different populations are included. It is unclear how well factor-analytic results from a single study accurately represent the "true" relationships among measures of different cognitive ability constructs.

One way to resolve this problem is to amass results from several studies conducted by several researchers and involving different populations. For the most part, research related to the development of the Kit of Factor-Referenced Cognitive Tests has utilized this approach. Over the years, the samples used to identify or replicate factors have included male Naval recruits, Army enlistees, college students, male and female eleventh- and twelfth- grade students, and sixth- and ninth-grade students. This research represents the most rigorous attempt to identify cognitive ability factors based upon samples differing in levels of education and ability.

Summary

In this subsection, we examined research programs designed to identify independent or distinct cognitive ability constructs. The first program described was Guilford's Structure-of-Intellect model, which proposed the existence of 120 or more independent cognitive ability constructs. This model contains three categories--operations, content, and products--into which the 120 cognitive ability constructs may be grouped. One of the most distinguishing features of the model is the group of divergent thinking factors designed to assess creative thinking processes such as originality, flexibility, and fluency. Studies investigating the validity of these measures have met with mixed results. That is, construct validity has been established for many of the measures. Predictive criterion-related validity, however, has been demonstrated for only a few of the divergent thinking measures.

The Structure-of-Intellect model has been criticized for the factor analysis procedures used to derive the independent cognitive ability constructs. It has been argued that the procedures Guilford used were too subjective and permitted confirmation of any hypothesized factors, including a randomly generated set of factors.

The second research program we examined was conducted by researchers at the Educational Testing Service, who pooled factor analysis results from studies conducted by different researchers utilizing subjects differing in ability and in education levels. From the accumulation of factor analysis data, these researchers constructed a battery of tests designed to measure independent cognitive ability constructs. The most recent battery, Kit of

Factor-Referenced Cognitive Tests, includes tests for 23 cognitive ability constructs (Ekstrom et al., 1976). In a 1979 monograph these researchers concluded that, at present, only 11 of the 23 cognitive ability constructs may be considered distinct or independent.

Our goal in reviewing these two research programs was to specify an initial structure of the cognitive abilities domain. A requirement of our cognitive ability taxonomy is an established linkage between measures of cognitive ability constructs and measures of training or job performance outcomes. This information can be used to evaluate the contribution that measures of each ability construct may make in selecting and classifying Army enlisted personnel. Research conducted and summarized by Guilford (Guilford & Hoepfner, 1971) and by Educational Testing Service researchers provided information to design a preliminary cognitive ability taxonomy. These research programs did not, however, attempt to establish a linkage between performance in measures of cognitive ability constructs and performance in applied settings. Therefore, in the next subsection we examine the content of several multi-aptitude test batteries employed for applied purposes and summarize validity data for these test batteries.

FROM THEORETICAL TO PRACTICAL APPLICATIONS

Description of Four Multi-Aptitude Test Batteries

Two major research projects described above were designed to explore the number of cognitive ability constructs, and represent an accumulation of well over 20 years of data. In both projects, data were analyzed to confirm or disconfirm the existence of independent or distinct cognitive ability constructs. For applied purposes, however, linkages between confirmed cognitive ability constructs and job performance constructs are yet to be established. At present it is unclear how the cognitive ability taxonomies generated from the Guilford and ETS research may be used to predict success in educational or occupational settings.

An alternative approach to the practical application question involves examining the types of cognitive abilities that are currently assessed for educational and occupational prediction purposes. Below we provide a description of two multi-aptitude test batteries used to predict academic performance, and two used to predict job performance. We review the procedures followed to develop each battery, define the cognitive abilities measured, and summarize psychometric information (e.g., reliability, validity) related to the effectiveness of each battery and corresponding subtests.

In the educational realm we examine Thurstone's (now Scientific Research Associates') Tests of Primary Mental Abilities, and the Differential Aptitude Tests (DAT). In industrial settings two of the more widely used batteries--the Flanagan Industrial Tests (FIT) and the Employee Aptitude Survey (EAS)--are described. The batteries we have chosen to explore were selected from among the many available for the following reasons: All four are widely used, and all are written, objective, and machine-scorable, which makes their use more convenient. Standardized instructions for administration are provided in test manuals; instructions are clear and easy to follow. Most of the subtests from these batteries have relatively short time limits. Reliability, validity, and other types of empirical data are available for each battery, as described below.

The Primary Mental Abilities (PMA) was normed on a large (N = 32,393) sample of individuals, ranging in age from four to 20. Descriptions of this sample, including age and grade distribution, appeared in the PMA Technical Report (Science Research Associates, 1965). Extensive reliability and validity data were also provided. Validity coefficients have been computed using both grade-point averages and other tests (e.g., Kuhlmann-Anerson test, Iowa Tests of Basic Skills) as criteria.

In the <u>Fifth Mental Measurements Yearbook</u>, the PMA received positive comments from both reviewers. Frederiksen (1959) concluded that the tests in the battery are theoretically sound and well constructed, while Kurtz (1959) pointed out that the battery is objective, is easy to administer, and has high face validity. Kurtz also agreed that the theoretical basis of the PMA is excellent.

Norms for the Differential Aptitude Tests (DAT) are based on a sample of more than 62,900 boys and girls in grades eight through 12. The DAT manual (Bennett, Seashore, & Wesman, 1973) is exhaustive, providing information on topics such as interpretation of individual profiles, principles followed in the development of the tests, and equivalence of alternate test forms. Reliability data are reported based on the results of testing 6,000 students, and the types of reliability calculated are both alternate forms and odd-even.

Extensive validity data, using large numbers of subjects, are also reported in the DAT Manual. The criteria used are course grades in English and literature, mathematics, science, social studies and history, business and business skills, and miscellaneous courses, including those at vocational high schools. Linn (1978a), in the <u>Eighth Mental Measurements Yearbook</u>, praised the DAT for its comprehensiveness and clarity, as well as for the manual's well-documented validities. He also noted that the development of the battery and the rationale behind its use are clearly articulated, and that the normative sample was chosen with care. Hanna (1978) was impressed with the "superb" format, clear directions, and good art work used in the DAT. Other reviewers have named the DAT as the best available instrument of its kind (Quereshi, 1972).

Because of the large amount of research that has been conducted on the PMA and the DAT, they represent valid educational aptitude batteries. Their counterparts in occupational settings are the Flanagan Industrial Tests (FIT) and the Employee Aptitude Survey (EAS). As in the case of the educational tests, research supporting these batteries provides sufficient justification for their discussion here. For example, the FIT manual (Flanagan, 1965) provides reliability data in the form of correlations with the Flanagan Aptitude Classification Tests (FACT). Subtest intercorrelations with the FACT, and with other FIT subtests, are reported. Norms for the FIT are based on high school students, college freshman (male), and industrial worker samples totaling 12,334. Validities are based on grade-point averages of college freshmen (N=701), and performance rankings of employees in a particular job category by their immediate supervisor (N=8284). The FIT validity studies examined a wide variety of job titles including samples of workers in clerical, maintenance, electronics, and heavy equipment operator positions (see Table 4, p. 39).

Adcock (1972), in reviewing the FIT in the Seventh Mental Measurements Yearbook, indicated it is a worthy and valuable tool for vocational selection. Horn (1972) generally agreed, and pointed out the usefulness of this battery for situations in which employers feel the need to tailor tests to their own local standards. A later reviewer (Herman, 1978) noted the practical benefits of the battery, such as convenient administration and scoring, and the ease of assembling smaller batteries for special purposes. Finally, MacKinney (1978) was impressed with the sizable amount of validity information available for the FIT.

The EAS has also been well researched. Its manual (Ruch & Ruch, 1963, 1980) reports alternate form and/or test-retest reliability estimates for each subtest. These estimates are based on samples ranging in size from 853 to 1,782. It is easy to apply applicants' scores for a particular test directly to the industrial setting, since the norms provided in the manual are categorized by job type; norms for 57 jobs range from secretary to industrial engineer, from chemist to clerk. Tables are also available providing norms for more general populations, such as male or female college students.

Scores on the EAS have been correlated with other aptitude tests; these are reported for the PMA, the Bennett Test of Mechanical Comprehension, the Otis Employment Test, the DAT, the California Test of Mental Maturity, the Minnesota Clerical Test, and the Cooperative School and College Ability Tests. A final reason for including the EAS for discussion is that validity coefficients have been computed for a wide variety of job groups, using industrial rather than educational criteria. Most often the criteria included supervisory ratings, but in some cases hired/not-hired status (after a trial period) or grades in training courses were used. Jobs for which validity coefficients were computed have been categorized into five major groups; clerical, sales, management and supervisory, skilled and semi-skilled, and technical.

Wallace (1959), in reviewing the EAS for the Fifth Mental Measurements Yearbook, concluded that the battery is well thought out and well constructed. He also remarked on the uniform excellence of the general format, administration instructions, and scoring keys. In the Sixth Mental Measurements Yearbook, Ross (1965) concurred with Wallace, and recommended the use of the EAS. In Taylor's (1965) review of the EAS technical manual, he stated that he was favorably impressed with the battery, and suggested that in preparing similar manuals for other batteries, researchers would do well to use this one as a model.

In sum, the batteries described in this subsection--the Primary Mental Abilities, Differential Aptitude Tests, Flanagan Industrial Tests, and Employee Aptitude Survey--were chosen for discussion because of their administrative convenience and because of the large amount of research that has been conducted with each of them.

Summary of Psychometric Data for the Four Batteries

Primary Mental Abilities (PMA). The reader will recall that Thurstone identified the "primary mental abilities" through orthogonal factor analyses. His next step was to develop or identify a test to measure each factor. Results from early research led him to conclude that some cognitive functions have a primary factor in common, distinguishing them from other cognitive functions, and thereby yielding groups of functions with different common primary factors. Thurstone collected a set of tests to measure these different cognitive function areas, resulting in the Tests of Primary Mental Abilities (PMA).

The most recent revision of the PMA includes tests to measure five abilities deemed to be most important in school work; forms of the PMA have been developed for use in grades K-12. Admittedly, these five abilities do not represent all of the factors that have been identified through research in the field. The five factors of intelligence measured by the PMA are defined as follows (Science Research Associates, 1965):

1. Verbal Meaning - the ability to understand ideas expressed in words.

2. Number Facility - the ability to work with numbers, to handle simple quantitative problems rapidly and accurately, and to understand and recognize quantitative differences.

3. Reasoning - the ability to solve logical problems.

4. Perceptual Speed - the ability to recognize likenesses and differences between objects or symbols quickly and accurately.

5. Spatial Relations - the ability to visualize objects and figures rotated in space and the relations between them.

Thurstone found that the general intelligence factor, or *g*, emerged as a second-order factor. Hence, he incorporated the option of using a single-quotient score derived from the PMA. This score is believed to provide a reliable estimate of intelligence, and should be comparable to Stanford-Binet or Wechsler Intelligence Scale for Children (WISC) scores.

Test-retest reliability estimates reported in the test manual (Science Research Associates, 1965) indicate that median values computed from 30 studies are quite high. Test-retest intervals in these studies range from one to four weeks. Median values for the five subtests and total score are as follows: Verbal, .89; Spatial .78; Number .82; Reasoning .83; Perceptual .81, and Total .91.

Subtests on the PMA have been evaluated against a criterion measure consisting of course grades (see Table 2). Subjects in the 26 samples included students in grades 2 through 12. The total median validity coefficient was .55.

Differential Aptitude Test (DAT). The PMA test ushered in the development of many multiple-aptitude batteries that yield a set of scores for an individual rather than (or sometimes in addition to) a single general ability score. By generating a profile of several broad aptitude areas, these batteries prove more useful in vocational counseling than do the global intelligence test scores; the latter provide no more than predictions of expected levels of attainment (Anastasi, 1964). It was with the intended purpose of aiding high-school counselors that the Differential Aptitude Test (DAT) battery was published by Bennett, Seashore, and Wesman of the Psychological Corporation in 1947. Bouchard (1978) pointed out that five of the eight aptitudes measured by the DAT overlap with those measured by the PMA. These are (1) Verbal Reasoning (VR), (2) Number Ability (NA), (3) Abstract Reasoning (AR), (4) Clerical Speed and Accuracy (CSA) (analogous to perceptual speed), and (5) Space Relations (SR). The definitions of these abilities are similar to those given by Thurstone in the PMA. The three additional tests included in the DAT are:

6. Mechanical Reasoning (MR) - the ability to learn and use the principles of operation and repair of complex devices.

7. Spelling (SP) - the ability to recognize misspelled words.

8. Language Usage (LU) - the ability to detect errors in grammar, punctuation, and capitalization.

The authors of the DAT recognize that the latter two subtests are more similar to achievement than to aptitude tests. Their rationale for including them in this aptitude battery is that they are believed to represent basic skills necessary in many educational and vocational pursuits. Together, the scores on these two tests estimate the ability to distinguish correct from incorrect English usage, an ability needed in many types of jobs. It is apparent that, unlike the PMA, the developers of the DAT focused on the measurement of complex abilities that are more directly related to jobs, rather than maintaining a strict emphasis on factorial "purity."

Reliability estimates computed for the DAT subtests were generated using the split-half internal consistency procedure. Samples for each subtest include about 250 subjects. Subtest reliability estimates range from .88 to .95. Values for each subtest and for total score are as follows: Verbal Reasoning, .95; Number Ability, .92; Abstract Reasoning, .94; Clerical Speed and Accuracy .89; Mechanical Reasoning, .88; Space Relations, .93; Spelling, .95; Language Usage .92; Total, .96.

The DAT Manual documents validity of each subtest against course grades for a large number of studies; the results are summarized in Table 3. The estimated validities are fairly high for traditional academic course grades

Table 2

Validities of Primary Mental Abilities (PMA) Subtests, Based Upon Course Grades

	Grade											
PMA Subtest	2	3	4	5	6	7	8	9	10	11	12	Median
Verbal Meaning	.40	.52	.57	.55	.62	.57	.62	.39	.41	.48	.37	.52
Spatial Relations	.30	.47	.46	.40	.29	.26	.48	.17	.03	.21	.03	.29
Number Facility	.56	.59	.63	.67	.59	.56	.66	.43	.32	.38	.25	.56
Reasoning	--	--	--	.70	.58	.59	.62	.48	.30	.52	.33	.55
Perceptual Speed	.46	.40	.43	.47	.48	.52	--	--	--	--	--	.46
Number of Samples	3	3	3	3	3	3	3	2	1	1	1	
Sample Size Range	77-87	45-91	56-79	62-70	53-69	44-77	62-101	77-205	194	206	219	

Note: Summarized from Primary Mental Abilities Technical Report, by Science Research Associates (1965). Chicago: Science Research Associates. Copyright by Science Research Associates in 1965. Reproduced by permission.

Table 3

Differential Aptitude Test (DAT) Validity Using Course Grades as Criteria[a]

Subtest	English Literature	Mathematics	Science	Social Studies & History	Business & Business Skills	Reading	Auto Mechanics	Carpentry	Drafting	Electronics	Machine Shop	Welding	Foreign Languages
VR	.46	.33	.45	.48	.36	.47	-.03	.22	.16	.21	.33	.13	.21
	(.52)	(.37)	(.47)	(.52)	(.42)	(.53)							(.43)
NA	.48	.53	.51	.54	.44	.53	.16	.01	.27	.30	.41	.11	.36
	(.52)	(.49)	(.51)	(.51)	(.56)	(.50)							(.49)
AR	.39	.33	.38	.37	.38	.38	.19	.12	.29	.13	.41	.24	.22
	(.38)	(.38)	(.43)	(.40)	(.38)	(.43)							(.34)
CSA	.15	.10	.10	.18	.20	.18	.07	-.15	.20	-.08	.14	.14	.10
	(.10)	(.13)	(.12)	(.11)	(.10)	(.11)							(-.01)
MR	.24	.20	.32	.25	.30	.31	.22	.03	.28	-.03	.20	.49	.06
	(.30)	(.27)	(.25)	(.33)	(.29)	(.25)							(.12)
SR	.31	.27	.34	.29	.30	.31	.24	.13	.47	-.06	.02	.00	.19
	(.34)	(.31)	(.29)	(.33)	(.28)	(.35)							(.17)
SP	.35	.25	.36	.40	.26	.47	-.15	-.09	.18	.26	.15	.01	.18
	(.40)	(.30)	(.38)	(.42)	(.39)	(.45)							(.36)
LU	.47	.35	.48	.48	.47	.53	.10	.02	.35	.10	.21	-.07	.33
	(.50)	(.37)	(.44)	(.51)	(.47)	(.52)							(.39)
VR & NA	.51	.48	.54	.56	.42	.52	.06	.18	.28	.28	.39	.14	.36
	(.56)	(.51)	(.53)	(.58)	(.56)	(.57)							(.51)
# of Studies	69	48	35	53	12	4	2	3	7	1	11	1	7
	(71)	(46)	(32)	(58)	(21)	(4)							(10)
N Range	31-298	26-255	27-251	25-256	28-251	88-203	40-42	45-52	25-56	117	31-46	29	25-46
	(30-287)	(27-233)	(26-216)	(24-233)	(28-226)	(57-187)							(25-64)

Note: Summarized from Manual for the Differential Aptitude Test by G. K. Bennett, H. G. Seashore, and A. G. Wesman (1973), New York: The Psychological Corporation. Copyright by the Psychological Corporation in 1973. Reproduced by permission.

[a]Numbers in parentheses refer to validity estimates for samples of females, provided when available.

(.20 to .58), with the exception of one subtest (Clerical Speed and Accuracy, CSA). For vocational-technical courses, the predictive validity of the DAT subtests is somewhat lower (-.15 to .47).

The Verbal Reasoning (VR) test has one of the highest median validities of the DAT tests--approximately .40. This test shows high correlations with grades in academically oriented courses such as English literature, social studies, and history. This is not unexpected, because this measure was designed to assess the ability to understand concepts framed in words. It lacks validity as a predictor of manual dexterity types of skills such as welding, drafting, and auto mechanics (median r = .13).

The Numerical Ability (NA) test was designed to test one's understanding of numerical relationships and facility in handling numerical concepts. When correlated with traditional course grades, it appears to be very good as a predictor. Validities range from .44 to .56, with a median value of .51.

Abstract Reasoning (AR) is a nonverbal measure of one's reasoning ability or ability to discover principles guiding change in geometric figures. It appears to be a generally valid scale for traditional courses, with coefficients ranging from .33 to .43, and correlations with vocational-technical courses also show acceptable values (median r = .21).

Clerical Speed and Accuracy (CSA) measures response speed in one's reactions to simple letter and number combinations; it was not designed to measure any intellectual component. Its validity coefficients are all low, ranging from -.15 to +.20. This is the lowest validity for any DAT subtest.

Mechanical Reasoning (MR) assesses one's understanding of the principles of common physical forces. Scores on this measure may be influenced by exposure to mechanical or shop courses. This measure shows generally lower validities than do the other DAT subtests (range -.03 to .49 with median r = .25), as well as the lowest retest reliability (median value = .88).

The Space Relations (SR) test requires mental manipulation of three-dimensional objects. It is distinguished from the abstract reasoning subtest in that the latter does not measure visual discrimination capacity. Validities range from -.06 to +.47 with a median value of .29.

The Spelling (SP) test is self-descriptive, and the manual notes that items were carefully chosen to be of equal difficulty. The Language Usage (LU) test measures one's ability to detect errors in grammar, punctuation, and capitalization. The test authors point out that the two tests correlate highly, and are measures of achievement rather than pure aptitude. Both appear to predict success in traditional course areas to a high degree, although spelling shows slightly lower predictive validity than does language usage. Validity estimates for the Spelling test range from -.15 to .47 with a median value of .25. Reported validities for Language Usage range from -.07 to .53 with a median value of .35.

When used together, the Verbal Reasoning and Numerical Ability subtests provide an estimate of general learning ability, according to the test authors. Its predictive validity for traditional courses ranges from .42 to .58 (median = .53).

Although the DAT has been widely used and often praised, the battery has received some criticism, which, for the most part, centers around the high subtest intercorrelations. Linn (1978a) suggested that these values indicate substantial redundancy and may, in fact, lower differential prediction. Quereshi (1972) also noted the high intercorrelations and pointed out that some combination of a subset of the tests would probably suffice in many cases. This leads to a related criticism, the lack of differential validity for different external criteria. Schutz (1965), in the Sixth Mental Measurements Yearbook, reported this conclusion as did Linn (1978a). Although extensive predictive validity exists for the DAT, differential validity of the subtests for the prediction of criteria has not been demonstrated. Bannatyne (1975) referred to the absence of any external validity results with regard to the DAT, and called it his "greatest disappointment."

Flanagan Industrial Tests (FIT). A test battery that has been widely used for selection in industry is the Flanagan Industrial Tests (FIT) battery. It is based on Flanagan's work during the 1940s with U.S. Air Force cadets. He found that training time could be considerably reduced by administering tests prior to assignment and using the results to place cadets in the job type for which their aptitude was greatest. This idea is easily generalizable to nonmilitary occupations, and at the end of the war Flanagan designed an aptitude battery for that purpose. It was called the Flanagan Aptitude Classification Test, or FACT. Its subtests were developed to measure distinct components of a job derived through job analyses. Since the various job functions were assumed to be separate and independent, the subtests were designed to measure distinct aptitudes.

The Flanagan Industrial Tests battery was developed from the FACT specifically for the purpose of selecting personnel for a wide variety of jobs. It is actually a short, speeded version of the FACT battery, designed exclusively to be used in adult populations. Each subtest of the FIT, like the FACT, measures a distinct, non-overlapping job element. Therefore, job applicants need be given only the subtests relevant to aspects of the job for which they are applying. This adds flexibility to the battery, as it can be used with a wider range of job types. Flanagan (1965) noted that an appropriate combination of subtests is a better predictor of performance than is a longer general ability test. However, no empirical evidence is available on this issue with regard to the FIT battery.

The 18 subtests of the FIT are:

1. Arithmetic - ability to work quickly and accurately with numbers (add, subtract, multiply, and divide).

2. Assembly - ability to visualize how an object would appear if it were assembled from a number of separate parts.

3. Components - ability to locate and identify parts of a whole.

4. Coordination - ability to coordinate hand and arm movements smoothly and accurately.

5. Electronics - ability to understand electrical/electronic principles and to analyze diagrams of electrical circuits.

6. Expression - knowledge of correct English; ability to communicate ideas verbally.

7. Ingenuity - creative/inventive skill; ability to devise ingenious procedures, equipment, or presentations.

8. Inspection - ability to detect flaws in a series of articles quickly and accurately.

9. Judgment and Comprehension - ability to understand what is read, to reason logically, and to use good judgment in interpretation.

10. Mathematics and Reasoning - understanding of basic math concepts, and translation of ideas/operations into brief mathematical notations.

11. Mechanics - understanding of mechanical principles.

12. Memory - ability to learn and recall a term associated with an unfamiliar one.

13. Patterns - precise and accurate perception and reproduction of simple pattern outlines.

14. Planning - ability to plan, organize, and schedule.

15. Precision - ability to do precision work with small objects, requiring speed and accuracy in making appropriate finger movements.

16. Scales - ability to read scales, graphs, and charts quickly and accurately.

17. Tables - ability to read tables quickly and accurately.

18. Vocabulary - knowledge of words.

As is apparent, Coordination and Precision are not strictly cognitive abilities subtests; they involve psychomotor skill. Flanagan chose these 18 particular subtests because they represented complex abilities needed in many industrial jobs. They were designed to measure requirements common to various jobs, but may be used in different combinations for testing the unique abilities needed for a particular job. In addition, the tests may be combined in the following ways to yield other ability estimates:

Ability Measured	Tests Used
General Intelligence	Judgment and Comprehension, Mathematics and Reasoning, and Vocabulary. Also, Expression, Ingenuity, and Scales.
Verbal	Vocabulary and Expression.
Quantitative	Mathematics and Reasoning, and Scales.

Reliability estimates for the 18 FIT subtests range from .28 to .79, with a median value of .56. It should be noted that these estimates are correlations of the FIT with the FACT, and are valid estimates of the reliability of the FIT only to the extent that the FACT is a reliable battery. The reliability of the FACT, then, serves as an upper bound to the FIT.

Table 4 shows the validities of the FIT subtests against two types of criterion measures. The first consists of freshmen grade-point average for university male students. Using grades as the criterion, median validity estimates for the FIT subtests range from a median low of -.03 to a median high of .22. The median value across all subtests is .13. Using job performance ratings as the criterion, the median validity estimates range from .00 to .29, with a median of the medians equal to .16.

Employee Aptitude Survey (EAS). The last widely used occupational assessment battery to be discussed in this section is the Employee Aptitude Survey (EAS), published by Floyd and William Ruch of Psychological Services in 1963 and 1980. Ruch and Ruch (1980) traced its development both to the results of Thurstone's factor analyses, and to a group of predictive validity studies of other aptitude areas. Although Thurstone's primary mental abilities had statistical backing through factor analysis, not much applied work had been conducted using them, and hence the primary ability measures lacked empirical validity. Combining those factors with the results of validity research on other aptitudes led to the development of the EAS.

The ten EAS tests are:

1. Verbal Comprehension - ability to use words in oral and written communication and in planning.

2. Numerical Ability - skill in the four fundamental operations of addition, subtraction, multiplication, and division.

3. Visual Pursuit - ability to visually track a line from its finishing point, when it is embedded in other lines.

4. Visual Speed and Accuracy - ability to quickly and accurately determine whether a pair of numbers are the same or different.

Table 4

Median Validities of Flanagan Industrial Tests (FIT) Subtests

Subtest	Criterion Measure: Grades[a]	Job Performance	Number of Samples[b,c]
Arithmetic	.20	.19	17
Assembly	.03	.17	11
Components	.00	.11	8
Coordination	-.02	.11	13
Electronics	.10	.22	3
Expression	.26	.13	5
Ingenuity	.13	.18	10
Inspection	-.05	.00	17
Judgment and Comprehension	.26	--	4
Mathematics and Reasoning	.39	--	4
Mechanics	-.01	.22	9
Memory	.12	.12	11
Patterns	.12	.20	8
Planning	.19	--	4
Precision	.11	.07	15
Scales	.16	.16	16
Tables	.17	.21	16
Vocabulary	.26	--	4

Note: Summarized from Flanagan Industrial Tests Manual by J. C. Flanagan (1965). Chicago: Science Research Associates. Copyright by Science Research Associates in 1965. Reproduced by permission.

[a]Four samples; sizes range from 69 to 362.
[b]Sample sizes range from 74 to 390.
[c]Job Performance validities computed for the following occupational job types:

Assembler
Carpenter
Claims Auditor
Claims Examiner
Clerk (various industries)
Drafter
Electrician
Electronic Technician
Freight Car Repairer
Heavy Equipment Operator
Machinist
Maintenance Mechanic
Packer
Plumber
Refinery Operator
Salesperson - Driver
Secretary - Stenographer
Subscriber - Relations Clerk
Telegrapher
Warehouse/Materials Handler
Yard Clerk

5. Space Visualization - ability to visualize what familiar blocks would look like if rotated in space.

6. Numerical Reasoning - ability to determine which numbers should follow in a given number series.

7. Verbal Reasoning - ability to recognize whether available facts support a conclusion.

8. Word Fluency - ability to generate lists of words beginning with a given letter.

9. Manual Speed and Accuracy - a psychomotor test.

10. Symbolic Reasoning - ability to evaluate symbolic relations.

Alternate-forms reliabilities range from .75 to .93 for the EAS subtests, with a median coefficient of .84. Validity estimates were calculated based upon various measures of training and job performance, including grade in training, supervisor ratings, and hired versus not-hired status at the end of a trial period. The validity coefficients have a median low value, across subtests, of .03 and a median high value of .70. The median of the subtest median values is .30. Table 5 summarizes the available validity information for the EAS.

The PMA, DAT, FIT, and EAS each have been widely utilized for a two-fold purpose. First, they have been used for vocational guidance, especially when administered to high school students. By examining the scores on subtests of multi-aptitude batteries, counselors are able to inform young people whether they have the abilities required to do well in various careers in which they might be interested. The second purpose of these batteries is for industrial personnel selection. By administering the tests to job applicants, people in charge of hiring for their company are in a better position to make good choices. Decisions on selection of applicants best suited to the job requirements are facilitated by the additional information provided by the batteries.

<u>Summary</u>

Descriptions of four widely used multi-aptitude selection batteries were presented to highlight (a) the types of cognitive ability constructs currently assessed to predict educational training and work performance outcomes; (b) the procedures and rationale underlying the development of each battery; (c) psychometric characteristics of battery subtests (e.g., reliability and validity); and (d) critical evaluations of each battery.

Table 6 indicates the types of cognitive ability constructs and technical knowledge constructs that are measured by the subtests of these batteries. The eight cognitive ability constructs are based upon the present review of work conducted by Thurstone (1938a, 1938b), Guilford (Guilford & Hoepfner, 1971), and Ekstrom and associates (1979). Table 6 defines the constructs only

Table 5

Validities for the Employee Aptitude Survey (EAS) Based on Job Performance Criteria

Subtest	Validity Range	Median	Number of Samples
Verbal Comprehension	.00 to .81	.29	39
Numerical Ability	.16 to .70	.39	41
Visual Pursuit	.05 to .54	.25	17
Visual Speed and Accuracy	-.08 to .59	.26	28
Space Visualization	-.24 to .73	.30	31
Numerical Reasoning	.05 to .70	.39	38
Verbal Reasoning	.11 to .71	.33	26
Word Fluency	-.09 to .47	.22	18
Manual Speed and Accuracy	-.13 to .33	.15	15
Symbolic Reasoning	-.08 to .70	.38	21

Note: Summarized from Employee Aptitude Survey: Technical report, by F. L. Ruch and W. W. Ruch (1980). Los Angeles: Psychological Services. Copyright by Psychological Services in 1980. Reproduced by permission.

Table 6

Cognitive Ability and Technical Knowledge Constructs Measured by Four Widely Used Multi-Aptitude Batteries

Cognitive Ability Construct	Battery Subtest
Verbal Ability	PMA Verbal Meaning FIT Vocabulary EAS Verbal Comprehension
Numerical Ability	PMA Numerical Facility DAT Numerical Ability FIT Arithmetic EAS Numerical Ability
Reasoning	PMA Reasoning FIT Judgment and Comprehension FIT Mathematics and Reasoning EAS Numerical Reasoning EAS Verbal Reasoning EAS Symbolic Reasoning DAT Abstract Reasoning FIT Planning DAT Verbal Reasoning
Spatial Ability	PMA Spatial Relations DAT Space Relations FIT Assembly EAS Visual Pursuit EAS Space Visualization
Perceptual Speed and Accuracy	PMA Perceptual Speed DAT Clerical Speed and Accuracy FIT Inspection FIT Scales FIT Tables EAS Visual Speed and Accuracy
Memory	FIT Memory
Fluency	FIT Ingenuity EAS Word Fluency
Perception	FIT Components FIT Patterns
Technical Knowledge Construct	
Mechanical Aptitude	DAT Mechanical Reasoning FIT Mechanics
Electronics Knowledge	FIT Electronics
Language Mechanics	DAT Spelling DAT Language Usage FIT Expression

in terms of which battery subtests have been used to measure them. Later in this section, definitions in terms of abilities measured are provided.

At this point, however, it is important to note the different types of tests that can and have been used to measure a single construct. For example, reasoning ability has been measured by abstract, verbal, symbolic, and numerical reasoning subtests; spatial ability has been measured by tests of spatial visualization, space relations, visual pursuit, and ability to assemble parts. It appears from this list of tests and constructs that different item types as well as different tasks may be used to measure the same underlying ability. We conclude that some ability constructs, such as reasoning and spatial abilities, can be further defined by subfactors. To ensure a complete cognitive ability taxonomy, these subfactors should be identified and defined.

Multi-aptitude batteries designed to predict success in educational or training settings include the Primary Mental Abilities (PMA) battery containing five subtests and the Differential Aptitude Test (DAT) with eight subtests. Common to both batteries are measures of numerical ability, reasoning, perceptual speed and accuracy, and spatial ability. The DAT also includes technical knowledge tests such as mechanical aptitude, spelling, and language usage.

According to the data provided in the PMA and DAT test manuals, subtests appear highly reliable in measuring target cognitive ability constructs (e.g., test-retest estimates range from .78 to .89; median $\underline{r}$ = .82 for the PMA; internal consistency estimates range from .88 to .95 for the DAT). Subtest correlations with academic or training course grades indicate that PMA measures yield validities ranging from .30 to .50. Reported validities for DAT subtests, using traditional academic course grades as criteria, range from .20 to .58, with the exception of one subtest (Clerical Speed and Accuracy).

Two batteries widely used for occupational prediction purposes are the Flanagan Industrial Tests (FIT) and the Employee Aptitude Survey (EAS). Cognitive ability constructs common to both batteries include verbal ability, numerical ability, reasoning, perceptual speed and accuracy, spatial ability, and fluency. Both batteries contain measures of psychomotor ability constructs. In addition, the FIT contains measures of memory, mechanical aptitude, and electronics knowledge.

The FIT battery is designed to predict success in a wide range of industrial occupations. Estimated reliabilities based on correlations with the FACT range from .28 to .79 with a median value of .56. Subtest correlations with measures of job performance, such as ranking of subordinates' overall job success by the manager, indicate that these measures may be used to predict success in a variety of occupations (e.g., clerical, maintenance, electronics). Estimated validities across all subtests range from .00 to .22 with a median of .16.

The EAS battery (10 subtests) was designed to predict performance in both white- and blue-collar jobs. Job types for which validity data exist include clerical, technical, skilled, semi-skilled, unskilled, sales, executive, administrative, and supervisory. Subtest reliability estimates range from .75 to .93, with a median value of .84 for alternate-forms reliability estimates. When correlated with measures of job performance (e.g., hired vs. not-hired, supervisors' ratings, and grades in training courses), subtest validity estimates range from .03 to .70, with a median value of .30.

Several common cognitive ability constructs are assessed in each of the four multi-aptitude batteries. These include numerical ability, reasoning, spatial ability, and perceptual speed and accuracy. Additional cognitive ability constructs assessed in one or more batteries include verbal ability, memory, fluency, and perception. Two of the four batteries contain technical knowledge measures, such as mechanical aptitude, electronics knowledge, and language mechanics.

This review provides illustrative data about psychometric qualities of paper-and-pencil cognitive ability measures in current applications. From these data, it appears that paper-and-pencil cognitive ability constructs provide consistent information about one's standing on a target construct. Validity estimates for cognitive ability constructs are examined in greater detail later in this report, but the data presented here indicate that these types of measures have been linked to potential for success in educational and occupational settings.

SECTION SUMMARY AND CONCLUSIONS

In this section we focused on establishing a cognitive ability taxonomy to help structure and summarize information from a review of the cognitive abilities literature. To understand the cognitive ability domain, we reviewed the history of intelligence theory and measurement from very early times to the present, and examined theories and tools designed to study and measure intelligence. Trends in intelligence theory indicate that the structure of intellect may be viewed as representing one of four models: two-factor, multiple-factor, facet, or hierarchical. Although there is little agreement about the structure of intelligence, it appears most theorists agree that intelligence comprises a number of different abilities.

Data from research designed to isolate independent cognitive ability factors comprising intelligence (Guilford & Hoepfner, 1971; and Ekstrom et al., 1979) provided information about measuring a variety of cognitive ability constructs and about the relationships among different cognitive abilities. Evidence systematically linking these cognitive ability factors to measures of work performance, however, is not available from either project.

Various types of cognitive abilities currently measured for educational and occupational selection purposes in four multi-aptitude batteries were described. Information about the nature of intelligence and measurement of cognitive abilities reported by several researchers (e.g., Thurstone, 1938a, 1938b; Guilford & Hoepfner, 1971; Ekstrom et al., 1979) was utilized to identify eight broad cognitive ability constructs: verbal ability, numerical ability, spatial ability, reasoning, memory, fluency, perceptual speed and accuracy, and perception. Subtests of the four batteries were then classified as measures of one of the eight cognitive ability constructs.

Another construct measured by two of the four batteries involves mechanical aptitude, which we classified as technical knowledge because it appears to measure knowledge acquired through experience with mechanical objects. The construct also appears to assess a combination of abilities such as spatial ability, reasoning, and perceptual speed and accuracy (Anastasi, 1976). Data indicate that measures of this construct can be used to predict performance in training or in educational settings (median $\underline{r}$ = .23) and performance in occupational settings (median $\underline{r}$ = .25) for a wide range of job types. Because this construct appears useful in predicting performance outcomes in a wide variety of settings, we chose to include it in the cognitive ability taxonomy and to study it carefully in our review of the literature. Other constructs involving technical knowledge (e.g., electronics knowledge) were omitted from the taxonomy because they appear to be targeted toward only a few specific occupations or job types.

The cognitive ability taxonomy, then, was designed to include ability constructs that have potential for predicting performance in a wide variety of training and occupational settings. Constructing our taxonomy involved incorporating information obtained from theories of the nature of intelligence, research exploring the number of independent cognitive abilities, and measures currently linked with training or occupational performance outcomes. Three goals in designing this taxonomy were parsimony, comprehensiveness, and generality. In other words, the taxonomy would allow us to summarize validity data gleaned from a review of the literature, using as few constructs as possible while still representing the entire domain.

As indicated previously, different types of tests may be used to assess one's standing on a particular cognitive ability (e.g., reasoning). These tests may involve different tasks, such as reasoning using verbal material or reasoning with figures, numbers, or symbols. Different tests, then, may be used to assess different components of the same target ability construct. To ensure that all aspects of each cognitive ability construct are represented, we have identified ability subfactors, where appropriate, from research designed to examine the nature and structure of intelligence. Finally, we elected to include constructs that have potential for predicting success in a wide variety of occupations, but omitted technical knowledge constructs that appear useful for predicting success in only a narrow occupational range. The final cognitive ability taxonomy contains nine broad cognitive ability factors: (1) Verbal Ability, (2) Numerical Ability, (3) Spatial Ability, (4) Reasoning, (5) Perceptual Speed and Accuracy, (6) Memory, (7) Fluency, (8) Perception, and (9) Mechanical Aptitude. Further, for six of the nine ability constructs, subfactors have been identified. Table 7 lists and defines the nine broad cognitive ability constructs and lists their subfactors.

Table 7

Cognitive Ability Taxonomy: Factor and Subfactor Definitions

Cognitive Ability Factor/Subfactor	Definition
1. Verbal Ability	Ability to understand the English language.
a. Verbal Comprehension	knowledge of the meaning of words.
b. Reading Comprehension	ability to read and understand written material.
2. Number/Mathematical Facility	Ability to solve simple or complex mathematical problems.
a. Numerical Computation	speed and accuracy in performing simple arithmetic operations such as addition, subtraction, multiplication, and division.
b. Use of Formulations and Number Problems	ability to use algebraic equations to solve number problems.
3. Spatial Ability	Ability to visualize or rotate objects and figures in space.
a. Space Visualization	ability to visually manipulate or transform the components of a two- or three-dimensional figure to see how things would look under altered conditions.
b. Two-Dimensional Mental Rotation	ability to identify a two-dimensional figure when seen at different angular orientations.
c. Three-Dimensional Mental Rotation	ability to identify a three-dimensional object projected on a two-dimensional plane, when seen at different angular orientations either within the picture plane or about the axis in depth.
d. Spatial Scanning	ability to visually survey a complex field to find a particular configuration representing a pathway through a field.

(Continued)

Table 7 (Continued)

Cognitive Ability Taxonomy: Factor and Subfactor Definitions

Cognitive Ability Factor/Subfactor	Definition
4. Reasoning	Ability to discover a rule or principle and apply it in solving a problem.
a. Inductive Reasoning	ability to form and apply hypotheses that fit a set of data.
b. Deductive Reasoning	ability to use logic and judgment in drawing conclusions from available information.
c. Analogical Reasoning	ability to identify the underlying principles governing relationships between parts of objects.
d. Figural Reasoning	ability to generate and apply hypotheses about principles governing relationships among several figures.
e. Word Problems	ability to select and organize relevant information to formulate solutions for mathematical problems.
5. Memory	Ability to recall previously learned information or concepts.
a. Associative or Rote Memory	ability to recall one part of a previously learned but unrelated item pair when the other part of the pair is presented.
b. Memory Span	ability to recall a number of distinct elements for immediate reproduction.

(Continued)

Table 7 (Continued)

Cognitive Ability Taxonomy: Factor and Subfactor Definitions

Cognitive Ability Factor/Subfactor	Definition
c. Visual Memory	ability to remember the configuration, location, or orientation of figural material.
6. Fluency	Ability to rapidly generate words or ideas related to target stimuli.
a. Associational Fluency	ability to rapidly produce words that share a given area of meaning or some other semantic property.
b. Expressional Fluency	ability to rapidly think of word groups or phrases.
c. Ideational Fluency	ability to write a number of ideas about a given topic or examples of a given class of objects.
d. Word Fluency	ability to produce words that fit one or more restrictions that are not relevant to the meaning of words.
7. Perception	Ability to perceive a figure or form which is only partially presented or which is embedded in another form.
a. Flexibility of Closure	ability to "hold" a given percept or configuration in mind so as to disembed it from other well-defined or complex material (Field Independence).
b. Speed of Closure	ability to identify objects or words given sketchy or partial information (Verbal and Figural Closure).

(Continued)

Table 7 (Continued)

Cognitive Ability Taxonomy: Factor and Subfactor Definitions

Cognitive Ability Factor/Subfactor	Definition
8. Perceptual Speed and Accuracy	Ability to perceive visual information quickly and accurately and to perform simple processing tasks with it (e.g., comparisons).
9. Mechanical Aptitude	Ability to perceive and understand the relationship of physical forces and mechanical elements in a prescribed situation.

This taxonomy was used to structure and summarize our review and evaluation of the literature reporting data for paper-and-pencil cognitive ability measures. Before presenting the literature review summary, we describe events that led to the development of cognitive ability measures for use in occupational settings which, in turn, led to the development of test batteries for selection and classification purposes.

SECTION III

CONSERVATION OF HUMAN TALENT: DEVELOPING AND VALIDATING SELECTION TOOLS

Sections III and IV focus on a similar theme, conserving human talent in occupational selection. In Section III, we trace events that influenced the use of measures of intelligence in the work setting. First, the history and purpose of the Employment Stabilization Research Institute (ESRI) are discussed. This institute was influential because it introduced the idea of using multi-aptitude batteries for vocational assessment. Researchers at ESRI were among the first to link job-related abilities to cognitive tests.

Following this, we move ahead in history to World War II. The second topic describes research involved in identifying the appropriate criterion measure to validate selection tests. The third topic focuses on test development activities during World War II. This includes a discussion of procedures used to develop and validate selection measures, expansion of the cognitive ability domain in terms of numbers of abilities measured, and development of selection and classification systems, all of which have had impact on the current military screening and classification battery. The final topic examines the impact of the changing work force during and following World War II.

THE GREAT DEPRESSION

The Employment Stabilization Research Institute (ESRI)

The U.S. Employment Service (USES) has developed a multi-aptitude battery, the General Aptitude Test Battery, that is used nationally in some capacity in almost all USES locations (P. Dersden, personal communication, June 12, 1989; Schmidt, 1988). Before describing the battery, it is appropriate to report on its development, beginning with work conducted at the Minnesota Employment Stabilization Research Institute (ESRI) around the time of the Great Depression.

The Institute was established in 1930 at the University of Minnesota and was concerned primarily with the two great economic and social problems at that time--unemployment and relief (Nelson, 1955). From its inception, an interdisciplinary approach was used. Three projects were conducted simultaneously, each from a different field: economics, psychology/ education, and personnel administration. Each will be described in turn.

Objectives for studying the economic aspects of unemployment in Minnesota were threefold. First, Project I was aimed at determining the impact of industrial change on the amount and type of unemployment. Second, based on the data obtained, it sought to identify needs for vocational training and guidance. Finally, the project assessed possible changes that could be made in the organization and management of business to help alleviate unemployment.

The second project, which was concerned with individual diagnosis and retraining, also served three purposes (Stevenson, 1931): "(1) testing various methods of diagnosing the vocational aptitudes of unemployed workers; (2) providing a cross-section of the basic re-education problems of the unemployed; and (3) demonstrating methods of re-education and industrial rehabilitation of workers dislodged by industrial changes" (p. 15).

The personnel administration division of the ESRI (Project III) used public employment agencies to test the findings of the first two projects. Agencies serving as "testing grounds" were located in Minneapolis, St. Paul, and Duluth, Minnesota. A schematic representation of the ESRI organization chart is provided in Figure 1.

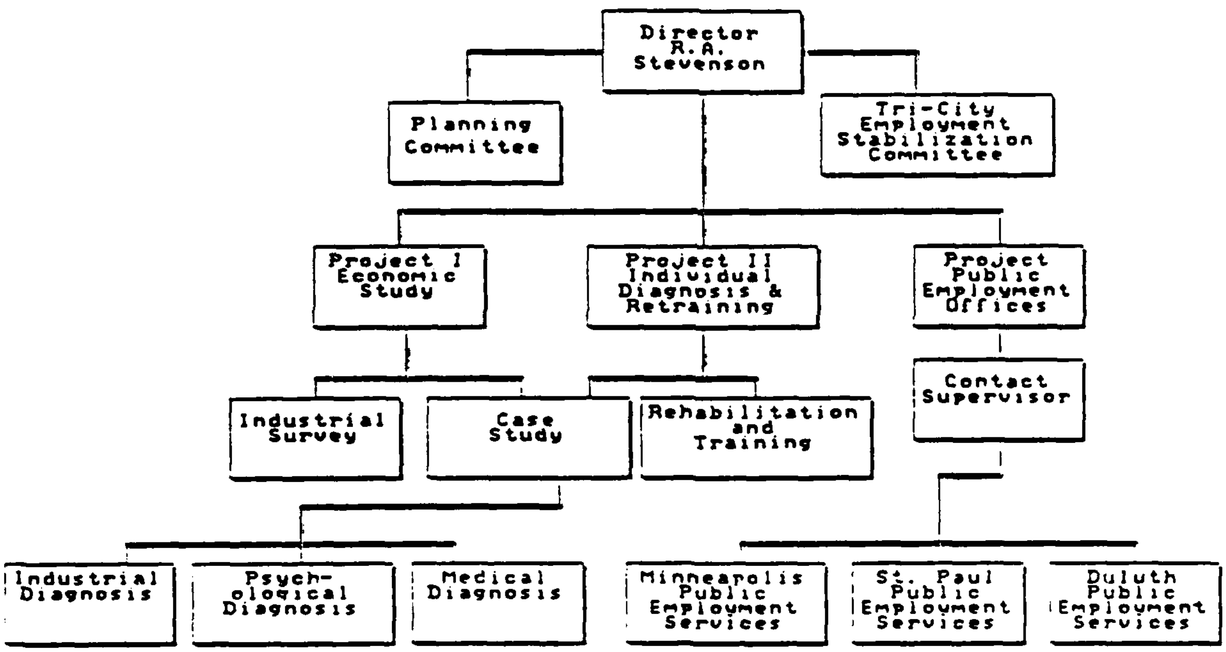


Figure 1. Organization of the Employment Stabilization Research Institute, Minnesota (summarized from Stevenson, 1931)

The second project, individual diagnosis and retraining, is the most relevant to this report, so it will be described in detail. To individualize the employment stabilization program, 4,000 unemployed persons in Minnesota were identified and classified on two dimensions: the cause of the individual's unemployed status, and the individual's actual or potential industrial usefulness. The second dimension enabled researchers to compute statistics to determine the proportions of unemployed persons (a) who were unfit for employment (due to either mental or physical incapacities), (b) who needed retraining prior to job placement, and (c) who were readily available for employment if the appropriate jobs were available.

The individual diagnosis involved three steps. First, each individual was interviewed in detail regarding his or her occupational and educational background. From the interview data an Occupational History Schedule was completed, to determine the individual's actual and potential occupational fitness. Interview statements were verified, primarily through checking school and social agency records.

The second step involved vocational testing in relation to occupational specifications. Stevenson (1931) pointed out that the rationale for these procedures was based upon the theory that groups of occupations have varying requirements in terms of interests and aptitudes, and that individuals' interests and aptitudes can be reliably tested and then matched to occupations. The abilities or characteristics assessed in the ESRI program were:

1. Educational Status (Grade)
2. Educational Ability (Academic Intelligence)
3. Clerical Aptitude
4. Manual Dexterity
5. Mechanical Aptitude
6. Strength of Hands, Back, and Legs
7. Vocational Interests
8. Trade Skill Proficiency
9. Personality Traits
10. Sensory Acuity

The last phase of diagnosis involved a complete physical and medical examination, emphasizing factors that might restrain an individual's work ability.

Project II was concerned with training program research. Its five objectives were:

1. Determining which individual differences are predictive of success in training.

2. Determining the predictive validity of the tests employed in the individual diagnosis phase of this project.

3. Identifying related types of jobs for persons who had been employed in now-obsolete jobs.

4. Developing and testing new training methods.

5. Helping individuals adapt to the work force by identifying their strongest aptitude areas and by training them (Stevenson, 1931).

The individual diagnosis and retraining projects at ESRI gave rise to the Adjustment Service. This agency was established in New York City in 1933 to provide vocational guidance for unemployed adults (Paterson & Yoder, 1955). A chain reaction started, leading to the creation of vocational guidance services in most schools and social agencies, and in the Veterans Administration.

The work conducted at the Employment Stabilization Research Institute is pertinent to this review of the literature on selection research because it was one of the first large programs utilizing a battery of selection devices. Many procedures which were developed later were based upon ESRI's Occupational Analysis approach--its ideas, principles, and, in some cases, the actual tests used in the analyses. For example, the major subtest used in the ESRI battery was a measure of general academic intelligence or verbal ability (the Pressey Senior Classification Test and Senior Verification Test). Today, almost all selection test batteries include some measure of general intelligence. The types of test items used in both Pressey tests are still in widespread use today; both tests included items of four types--opposites, information, practical arithmetic, and practical judgment (Paterson & Darley, 1936).

In addition to the general aptitude test, which was said to form the backbone of the battery, ESRI employed one clerical test and two tests of mechanical ability. The clerical test, the Minnesota Vocational Test for Clerical Workers, measures the quickness and accuracy with which one can perceive similarities and differences between pairs of numbers and between pairs of names (Paterson & Darley). This test appears similar to the perceptual speed factor tested in Thurstone's Primary Mental Abilities (PMA) battery, the clerical speed and accuracy aptitude of the Differential Aptitude Tests (DAT), and the visual speed and accuracy ability of the Employee Aptitude Survey (EAS).

The mechanical aptitude tests included in the ESRI Occupational Analysis were the Minnesota Mechanical Assembly Test and the Minnesota Spatial Relations Test. The assembly test requires the examinee to assemble a variety of mechanical devices, given the necessary parts. Tests used today for the purposes of selection and placement more frequently use written items with multiple-choice responses. These items show drawings of the parts of various objects, and require the examinee to select a drawing, from among a set of four or five alternatives, that most closely represents what the parts would look like if assembled. This format is more conducive to quick and easy machine-scorable group testing, and offers the additional benefit of eliminating some element of psychomotor skill from the test score. An example of a modern version of a test of this type is the Flanagan Industrial Tests (FIT) Assembly Test.

Researchers at ESRI considered the Minnesota Spatial Relations Test to be a test of mechanical aptitude because they believed that occupations such as auto mechanics, woodwork, sheet metal work, and complicated skilled trades required a large component of spatial ability (Paterson & Darley, 1936). This test consists of four boards with holes of odd shapes and sizes in which the examinee must place correctly shaped pieces. Because of the costs involved in administering this type of test, it is now rarely used in industrial settings. Different types of spatial abilities tests are, however, frequently included in selection batteries used today. As noted above, the PMA includes a measure of spatial relations, the DAT has a space reasoning subtest, and the EAS has a space visualization test. Like the more modern forms of assembly tests, these spatial relations tests are written and thus the component of psychomotor ability has been removed from the scores.

The ESRI researchers recognized the importance of personal characteristics other than intellectual aptitude for predicting job suitability. For this reason, their test battery included measures of dexterity (the O'Connor Finger Dexterity Test, the O'Connor Tweezer Dexterity Test, and the Minnesota Manual Dexterity Test) and of interests (the Strong Vocational Interest Blank for Men, and Manson's Womens' Occupational Interest Blank). Today, comprehensive batteries also include similar tests of physical or psychomotor skill, and interest inventories.

Development of the General Aptitude Test Battery (GATB)

The General Aptitude Test Battery (GATB) was developed from research conducted by the Occupational Analysis Division of the U.S. Employment Service, but the idea and principles underlying the product began at ESRI. First published in 1947, the battery was designed for two purposes: to measure individual aptitudes that have been found to be important in a wide variety of occupations, and to establish norms regarding these aptitudes so that comparisons could be made between an individual's profile and that of different job types.

According to the 1970 GATB manual (Department of Labor), subtests were selected using two criteria: (a) internal or factorial validity (size of factor loading across the different studies), and (b) external or practical validity (from occupational validation studies). These criteria resulted in the selection of 12 written tests and four tests requiring the use of apparatus. Based upon further study four of the written tests were eliminated. The current battery contains eight written and four apparatus tests. The tests, along with the nine aptitudes they are purported to measure, are listed in Table 8.

Scores on the GATB are given in the form of the nine aptitude scores. Originally there were 11 factors, but two (Aiming and Logic) failed to replicate. The initial validation of the GATB was conducted using nine different samples of young people, mostly teenagers. These individuals were either applicants for defense training courses or trainees enrolled in Vocational Education National Defense Training courses, representing 13 different geographic locations. The total sample included 2,156 subjects (N range = 99-1079) who completed from 15 to 29 tests.

Altogether, 44 tests, plus the GATB, were administered. These tests were considered to be representative of the more than 100 tests developed by the USES prior to 1942. Data were processed by factor analyzing the test intercorrelation matrix derived from each group, using Thurstone's

Table 8

Aptitude Factors Assessed by the General Aptitude Test Battery (GATB)

Aptitude Factor	Test	
G - Intelligence	Part 3 - Three-dimensional Space Part 4 - Vocabulary Part 6 - Arithmetic Reasoning	
V - Verbal Aptitude	Part 4 - Vocabulary	
N - Numerical Aptitude	Part 2 - Computation Part 6 - Arithmetic Reasoning	
S - Spatial Aptitude	Part 3 - Three-dimensional Space	
P - Form Perception	Part 5 - Tool Matching Part 7 - Form Matching	
Q - Clerical Perception	Part 1 - Name Comparison	
T - Coordination	Part 8 - Mark Making	
F - Finger Dexterity	Part 11 - Assemble Part 12 - Disassemble	Tests requiring apparatus
M - Manual Dexterity	Part 9 - Place Part 10 - Turn	

Note: From the General Aptitude Test Battery Manual, Section IIA, by the Department of Labor (1980).

multiple-factor analysis technique. Resulting factors were then orthogonally rotated using the simple structure criterion.

Each Occupational Aptitude Pattern (OAP) consists of minimum scores for the most significant aptitudes (only two to four are chosen) for the group of occupations represented by that particular OAP. Critical aptitudes were defined as those in which workers in a given job excelled the general norms, as well as those with significant correlations with criteria of job success. According to the GATB manual, the cutting scores are set so that the bottom one-third of the distribution of workers in each job is excluded (Department of Labor, 1980).

Once an individual has completed the GATB, his or her test scores are expressed as an Individual Aptitude Profile, which is then compared to the various OAPs. Through a profile-matching process, the examinee can be informed as to how similar his or her profile is to that of people currently employed in different occupations. This makes the GATB a useful instrument in vocational counseling and guidance. In the matching process, the multiple cut-off method is used; that is, no total score is calculated. Therefore, an applicant must achieve the minimum cut-off score on <u>each</u> aptitude.

Because some aptitudes show major changes until around age 16 or grade 11 (Super & Crites, 1962), reliability (and, hence, predictability) is higher if the test is administered after that age. Fortunately, 16 is also the age at which most vocational guidance is needed.

Summary

In this section, we examined events during the Great Depression that helped to further selection test development. The primary event was the formation of the Employment Stabilization Research Institute (ESRI). Because one of the major goals of this organization was to identify employment opportunities for the many unemployed during the Great Depression, one branch of the institute developed a three-step process for individual diagnosis. This included: (a) obtaining background information; (b) administering a battery of tests such as general intelligence, mechanical aptitude, and clerical ability; and (c) conducting physical and medical examinations. Results from this research led to the development of one of the most widely used selection batteries, the General Aptitude Test Battery.

The GATB contains eight paper-and-pencil and four apparatus measures that when used alone or in combination provide information for general intelligence, five cognitive ability constructs, and three psychomotor constructs. Scores on these tests are used to identify person-job matches; this information is used in vocational counseling and guidance.

In sum, research conducted at ESRI was of major importance because the objective testing procedure utilized there established a standard for all future selection programs. Although some of the tests in use today are not identical in format to ESRI's original tests, most batteries are similar in general content to the one developed at ESRI.

WORLD WAR II: CONSTRUCTING MEASURES OF WORK PERFORMANCE

Pre-World War II

Vocational guidance information gathered in the years before World War II during the Employment Stability Research Study, and the procedures developed by the U.S. Employment Service to identify the aptitudes and abilities needed to perform in specific occupations, provided invaluable data for both the military and the industrial sectors at the onset of World War II. For example, results from numerous job analyses helped to identify the worker characteristics required for success in thousands of military and civilian occupations (Shartle & Dvorak, 1943). In addition, these procedures were used repeatedly throughout the war period to design selection and classification systems.

Also during the period following World War I, psychologists began emphasizing the need to demonstrate the effectiveness of selection measures. In other words, "it became a basic philosophy that scores on tests used to select workers for a given job must be shown to be related to the degree of success they achieved on that job" (Ghiselli, 1966, p. 6). Prior to that period, the validity of a test was established by its correlation with measures of similar constructs. For example, during World War I the Army Alpha and Army Beta were designed to provide, in a group-administered paper-and-pencil measure, the same type of information obtained from Binet's intelligence test or the Stanford-Binet. The criterion used to select subtests for inclusion in the Army Alpha and Beta was the correlation of each with scores on the Stanford-Binet (Yerkes, 1921). Not until after the war were scores on the Army Alpha linked with job performance or job training measures (Harrell & Churchill, 1941). Following World War I, it was more common to find researchers assessing the practical utility of selection measures by correlating test scores with scores on work performance measures (e.g., Anderson 1929; Schultz, 1936; Viteles, 1929). Not until World War II, however, were the procedures for identifying and evaluating a criterion measure fully explicated.

Criterion Development and Evaluation

During the later 1930s and in early 1940, as the United States was drawn closer to war, numerous programs were established to help prepare for the task of selecting and classifying military personnel. One program initiated by the Army Air Force in the spring of 1941 was the Aviation Psychology Program, which was established to develop measures to select and classify aircrew personnel (e.g., pilots, bombardiers, and navigators). Because winning the war was considered to be highly dependent on air power, great amounts of research time and personnel were allotted to this program.

It was in this program that one group of researchers established the methodology for criterion development. Thorndike (1947), in his summary of the problems encountered in the research program, noted that the criterion problem was the most fundamental and the most difficult problem to resolve. To address the criterion issue, Thorndike conceptualized the nature of the problem in terms of the types of criterion measures available. For example, he described and differentiated among three types of criterion job measures available at different points in time and identified factors one may use to evaluate these different types of measures--the immediate, the intermediate, and the ultimate criterion.

The immediate criterion is the measure that becomes available most quickly and directly. In terms of aircrew performance, an immediate criterion would consist of graduation versus elimination from the first pilot training course, primary pilot training.

The intermediate criterion becomes available at some later point. For pilots, these would include graduation versus elimination from later training courses, basic or advanced pilot training. They could consist of supervisory ratings in advanced pilot training or in theater combat operations. Both types of measures are only partial criteria because they do not fully represent the ultimate criterion. The goal in developing the intermediate criterion is to identify performance that closely represents or correlates highly with ultimate criterion measures. In theory, then, all intermediate criterion measures developed should correlate highly with each other.

The ultimate criterion represents the final goal of a particular type of selection or training program. In terms of aircrew performance, for bombardiers this would consist of dropping bombs with maximum precision under combat conditions; for a career gunner it would include the maximum possible number of hits upon attacking fighter planes. Thus, for military occupations, the ultimate criterion measure includes performance under combat conditions. These conditions generally involve unpredictable variables and require interaction among personnel, resulting in multiple and complex criterion measures. Quite often, the ultimate criterion is unavailable or difficult to measure.

Factors on which criterion measures may be evaluated include: (a) relevance--performance measures that require the same abilities, knowledge, and skills as those required in the performance of the ultimate criterion measure; (b) reliability--primarily a statistical measure with unreliability caused by intrinsic (inconsistent performance) and extrinsic (fluctuation in external conditions) factors; and (c) freedom from bias--assurance that the same standards are used to evaluate different subgroups. (Current industrial psychology textbooks include another evaluation factor--practicality, or the cost-related factors involved in measuring work performance behaviors.)

Finally, Thorndike described the measures that served as intermediate criterion measures of aircrew job performance. These are listed below along with examples of each from the AAF Aviation Psychology Program study:

- Job Knowledge Tests - printed proficiency tests asking examinees to compute values to determine position, altitude, fuel consumption, and so on.

- Simulated Job Samples Scored Objectively - tests measuring skill in tracking and framing an attacking fighter. All activity is recorded by a gun camera.

- Subjectively Scored Job Samples - performance in stripping and assembling the .50 caliber machine gun, scored by an observer.
- Rating Scales - ratings provided for an entire mission or a complete segment of training.
- Summary Performance Records - percentage of hits in fixed gunnery for fighter pilots.
- Summary Academic Grades - may include average grades from primary, basic, or advanced pilot training.
- Summary Ratings - routine efficiency ratings required on all officer personnel.

According to the validity data reported in the series of AAF Aviation Psychology Research program reports, most tests were evaluated using an immediate criterion measure, graduation/elimination from primary pilot training. Although intermediate criterion measures were later used to evaluate selection and classification tests, the ultimate criterion measure, combat performance, was still difficult to capture in its full scope even at a time when such data were, in theory, potentially available.

Development and use of criterion measures to validate selection and classification instruments was not limited to the AAF Aviation Psychology Program. The Army General Classification Test (AGCT) and trade test classification devices were also developed, and were revised using job-related criteria information collected during the war (Staff, Personnel Research Section, 1947). Psychologists conducting research on these selection and classification devices also noted difficulties in identifying and obtaining the appropriate criterion measures. In other words, the ultimate criterion measure, the behavior of soldiers under combat conditions in jobs for which they were trained, was difficult, if not impossible, to obtain. Thus, one of the most readily available performance measures, training course grades, often served as the criterion measure in selection test validation research. One major problem with this criterion measure, however, was the common practice of instructors passing all or nearly all soldiers in training.

Other criterion measures used to validate the AGCT test scores included non-combat job performance measures, such as the number of words per minute transmitted or received by a radiotelegraph operator. Although these measures of performance provided useful information about technical job knowledge, the correspondence between non-combat performance and performance in actual combat situations was unknown (Staff, Personnel Research Section, 1943b). It appears that problems related to identifying criterion measures that plagued researchers during World War II are virtually unchanged from those encountered today.

One useful psychological tool devised during this period was the critical requirement technique, which involves asking persons familiar with the job (e.g., trainers, supervisors, or job incumbents) to describe, in behavioral terms, examples of effective and ineffective job performance. Flanagan (1954) describes the ways in which this tool was used in the AAF

Aviation Psychology Program. These included: (a) identifying specific reasons for failure in pilot training; (b) identifying reasons for bombing mission failure; (c) isolating effective and ineffective examples of combat leadership; and (d) understanding problems related to flying while experiencing vertigo or acute disorientation.

The critical requirement, or critical incident technique as it was later termed, provided useful information for analyzing the critical components of jobs, developing tests to measure the required abilities and skills, designing training programs, assisting with human factors engineering (especially in cockpit design), and developing criterion measures of job performance. Further, this procedure set the stage for a later milestone in criterion development, namely, Smith and Kendall's (1965) demonstration of using the critical incident technique to develop performance appraisal rating forms. The resulting Behaviorally Anchored Rating Scales (BARS) provide raters with behavioral descriptions of the critical job components or dimensions and with behavioral effectiveness level anchors agreed upon by persons familiar with the job (e.g., supervisors and trainers).

Summary

Prior to World War II, researchers began exploring the linkages between performance on cognitive ability tests and performance in a work setting. Not until World War II, however, did the criterion for job performance receive great attention. During this period, military researchers formulated systematic procedures to validate experimental selection tests. Thorndike (1947) laid out a theory for developing and evaluating criterion measures of job performance. These procedures provide a basis for a criterion development methodology that is still in use today.

Flanagan (1954) reported that the Army Air Force used the critical requirement (or critical incident) technique to identify reasons for pilot failure. Results from the technique were then used to develop or improve training programs and to modify equipment. This information was also used to construct selection measures. That is, Army Air Force researchers examined reasons for failure on the job and then generated ideas about ability measures that might help to screen out persons likely to fail for those reasons. This research led to the development of literally hundreds of selection tests.

WORLD WAR II: ADVANCES IN PREDICTOR DEVELOPMENT AND IMPLEMENTATION

Research conducted during World War II is noted for producing numerous milestones in psychological assessment and tool development. Establishing a methodology for criterion development and using the critical incident technique for job analyses, selection, training, and criterion development purposes are two that have already been discussed.

In the area of selection and placement, other milestone events occurred. During the war hundreds of selection measures were developed and validated. Classification schemes involving the newly developed selection measures were designed to make efficient use of the individual's skills and abilities. To understand how the Army's selection and classification procedures evolved, we first examine the use of selection measures prior to World War II. Following this, we examine the selection and classification procedures developed during World War II and then briefly review the procedures designed for selection and classification purposes following the war and those currently used by the Army.

Initial Selection and Classification Measures

The Army Alpha. During World War I, Yerkes and his colleagues developed measures to aid in the selection of enlisted personnel. The Army Alpha and its nonverbal counterpart, the Army Beta, were designed to measure the ability to learn, to think quickly and accurately, to analyze the situation, to maintain a state of mental alertness, and to comprehend instructions (Yerkes, 1921). To utilize the information derived from the Army Alpha, individuals were assigned letter grades based on obtained test scores. Test score ranges and letter grades are listed in Table 9 along with the corresponding scores from the Stanford-Binet.

Table 9

Correspondence Between Army Alpha Test Scores and Stanford-Binet Test Scores

Letter Grade	Test Score Range[a]	Equivalent Stanford-Binet Score
A	140-212	18.0 - 19.5
B	110-139	16.5 - 17.9
C+	80-109	15.0 - 16.4
C	50-79	13.0 - 14.9
C-	30-49	11.0 - 12.9
D	15-29	9.5 - 10.9
E	0-14	0.0 - 9.4

Note: From Psychological Examining in the United States Army by R. M. Yerkes (1921), Memoirs of the National Academy of Sciences, Vol. XV.

[a]These values are based on raw scores summed across the eight tests included in the Army Alpha.

Interpretation of these letter grades is as follows: Grades A and B were typical of officers; Grade C was typical of privates; Grades D and E represented lower levels of intelligence (Yerkes, 1921). Application of test scores and letter grades for selection and classification decisions was at the discretion of organization commanders, who decided whether and to what extent to use test information. Basically, when Army Alpha and Army Beta test information was used, it aided in decisions related to: (a) selecting officers and NCOs, (b) identifying men for discharge, for labor battalions, or for special training battalions, (c) balancing or matching units by test score; and (d) identifying homogeneous training groups with respect to test scores. Thus, systematic testing of all incoming recruits using a group-administered paper-and-pencil measure for selection purposes was conducted during World War I. Systematic use of the selection tests for classification purposes was not, however, implemented during this period.

Following World War I, little research was conducted to learn how the Army could best make use of abilities identified in selection measures. In fact, during the period between 1918 and 1939, the Army continued to test recruits but made little use of psychological devices in selection and classification (Staff, Personnel Research Section, 1943a).

The Army General Classification Test (AGCT). When it became apparent that war was imminent, several agencies were established to expand the use of tests for selection and classification purposes. For example, during the spring of 1940, the Personnel Research Section was established in the Adjutant General's Office and Walter Bingham was named Chairman of the Committee of Classification of Military Psychology. Other members included C. C. Brigham, H. E. Garrett, L. L. Thurstone, L. J. O'Rourke, M. W. Richardson, and C. L. Shartle.

It was this committee that developed a classification test for the Army. The resulting test, the Army General Classification Test (AGCT), was designed to measure "general learning ability," and contained verbal, quantitative, and spatial ability items ordered in a spiral omnibus fashion. Examinees were given 40 minutes to complete a 150-item test. Raw scores[4] on the AGCT were converted to standard scores[5] which were used to assign enlistees to one of five categories. Category scores were then used to allocate men to different units. The categories, standard score ranges, and percentage of recruits in each category are listed in Table 10.

---

[4]The procedure for calculating the raw score included the number of correct responses minus one-third the number wrong.

[5]Standard scores were computed using the following formula: .82 raw score + 38.33, yielding a mean of 100 and a standard deviation of 20.

Table 10

Army General Classification Test: Category Scores, Standard Scores, and Percentage of Recruits[a]

Category	Standard Score Range	Percentage of Recruits
I	130 and above	6.0
II	110 to 129	26.7
III	90 to 109	30.3
IV	60 to 89	27.7
V	59 and below	9.3

Note: From The Army General Classification Test by the Staff, Personnel Research Section (1945).

[a]N = 8,293,879 recruits tested from 1940 to 1944.

In addition to the AGCT, other tests were developed by Personnel Research Section staff to assist in selecting and classifying Army personnel. For example, a minimum literacy test, visual classification test, and non-language test were developed to screen non-English speaking persons and persons of questionable ability. Special trade tests, such as Mechanical, Clerical, Radio Code Learning, and Automotive Information, were developed for classification purposes. Several tests, such as the Officer Candidate test and numerous Warrant Officer tests, were designed to identify potentially successful officers from among enlisted personnel. Additional batteries for special personnel or specialized occupations, including the Women's Classification Test and a battery for Combat Intelligence personnel, were developed (Staff, Personnel Research Section, 1943a).

All tests developed by the Personnel Research Section staff were constructed using the following procedures: (a) conducting an occupational analysis of a specialty field; (b) using information from technical experts, the technical literature, and other tests to develop test items; (c) conducting pilot tests of newly developed measures; (d) assessing the psychometric characteristics of the measures (e.g., reliability and validity); and (e) revising test items and standardizing test scores (Staff, Personnel Research Section, 1943b).

It is clear that the selection and classification measures developed during this period were constructed using comprehensive test development procedures still in use today. In other words, information about critical job requirements and input from technical experts or others familiar with the target job are used to determine the content for a particular selection measure. Following test development, pilot studies are conducted, and resulting test data are used to assess psychometric characteristics of the measure, such as test reliability and item difficulty levels. This information is used to revise the test, which is then validated against a criterion measure of job performance. As a result, these procedures or test development steps yield a psychometrically sound measure that is directly linked to the critical requirements of the target job or occupation.

During World War II, the Army's selection system included screening all enlistees or draftees for literacy. At the induction center, enlistees or draftees were asked to demonstrate reading and writing competencies at the fourth-grade level. Those who failed to meet the literacy requirements completed one or more of the following measures: a minimum literacy test, visual classification test, and two individual mental ability tests, Concrete Directions and Block Counting. Persons meeting the fourth-grade literacy requirements or passing one or more of the above tests were inducted into the Army (Uhlaner, 1952). At the Reception Center, inductees completed the AGCT[6]. Those obtaining low scores on the AGCT were asked to complete the non-language test; all others completed the Mechanical Aptitude, Radio Code Learning, or other trade tests. Recruits were then interviewed to determine educational level, job history, interests, hobbies, and previous military experience. In general, this information was used to classify those in Categories IV and V into Engineer, Infantry, and Signal Corps occupations. Additional tests (e.g., Officer Candidate Test) were administered to those in Categories I through III and these recruits were then assigned to specialist training.

In addition to the test and interview information, occupational classification was also based on quotas or the numbers required in each job. As the research staff notes, although the emphasis on filling quotas resulted in some misplacement of recruits, the primary objective was to ensure that all occupations were sufficiently staffed (Staff, Personnel Research Section, 1943b).

Regarding the construct validity of the AGCT, scores on it correlated fairly highly with other measures of general intelligence from that period. For example, AGCT scores correlated .83 with the Otis Test of Higher Mental Ability; with the American Council of Education Psychological Examination, AGCT scores yielded correlations ranging from .65 to .79; and with the Wells

[6]During World War II, the AGCT was used solely for classification purposes. In 1942, a shortened version of AGCT, R-1, was implemented to screen inductees with physical disabilities.

Revised Army Alpha,[7] the correlations range from .70 to .90 (Staff, Personnel Research Section, 1945, 1947). The AGCT yielded criterion-related validities ranging from .20 to .73 against training course grades in 30 technical specialty courses (Staff, Personnel Research Section, 1945).

Aviation Psychology Program of the Army Air Force

Cadet Qualifying Exam. Another research program, mentioned earlier, the Aviation Psychology Program, produced a wealth of information about ability constructs linked to measures of job performance. This program, directed by Dr. John Flanagan, was a large-scale effort by the Army Air Force to predict success in a narrow occupational group, air crew members. The thrust of the program was to rapidly and effectively identify

---

[7]To compare test score results obtained in World War I and World War II, a representative sample of World War II recruits completed the Wells Revised Army Alpha (N = 768). To reflect the mean educational difference between the two samples (8 years versus 10 years), the Army Alpha test scores for World War I sample were adjusted. Below are the corresponding percentile values for the WWI sample raw and adjusted scores and the WWII sample raw scores (Tuddenham, 1948).

Percentile	WWI Alpha Raw Score	WWI Weighted or Adjusted Score	WWII Wells Revised Alpha Raw Score
90	120	144	160
80	98	125	143
70	84	110	130
60	72	97	116
50	62	85	104
40	52	73	90
30	44	61	74
20	35	49	58
10	25	34	38

Even after adjustments were made for educational differences, the World War II sample scores are higher on the average than the World War I sample. This may have been due to the differences in the tests completed by the two samples and to differences in test-taking skills. Tuddenham also postulated that differences in health and nutrition might account for the higher scores in the World War II sample. This hypothesis seems questionable given that the World War II sample had been exposed to a lengthy period of depressed economic conditions. Humphreys (1986) suggests that even though the educational differences between the two groups on the average are small, these data indicate the influence that education can have on measured intelligence over long intervals (i.e., 24 years).

potentially successful candidates to serve as pilots, navigators, and bombardiers. Previously, from 1927 to 1942, recruits accepted into the Army Air Force were required to have two years of college education. To speed up the enlistment process, early in 1941, recruits with a high school degree and with passing scores on the AGCT, the Mechanical Aptitude Test, and a physics test were also allowed to enter the program.

Early in 1942, it was decided to establish a single set of entry requirements for all Army Air Force aircrew enlistees. Thus, educational level, AGCT, and mechanical aptitude and physics test score requirements were discarded; instead, potential candidates were required to obtain passing scores on the Aviation Cadet Qualifying Examination. This exam included measures of general vocabulary, reading comprehension, practical judgment, mathematics, current affairs in aviation, and mechanical comprehension[8].

Like the AGCT, the Cadet Qualifying Examination contained 150 items. Unlike the AGCT, however, this test was considered a power measure; examinees were given three hours to complete the test, but most completed it in under two hours. Approximately 33 to 50 percent of the examinees failed the exam and were dropped from further consideration in the program[9] (Flanagan, 1947).

<u>Classification Battery</u>. Following the initial selection process, a classification system was used to assign Army Air Force personnel to one of the aircrew positions. Briefly, this system was developed by first examining the reasons for failure in the pilot program and in navigator and bombardier training. Results from this investigation uncovered several ability and personal characteristic requirements common across the three aircrew positions and some unique to each[10]. Research units were established to develop and study the effectiveness of measures in each of

---

[8]The Aviation Cadet Qualifying Examination underwent 15 revisions during the war. Although the exams varied in length and in item content, subtests measuring verbal ability, current affairs, mechanical comprehension, mathematics, judgment, and interpretation of data appeared in nearly all versions. Subtests measuring perceptual abilities appeared only in the last four versions.

[9]Early versions of this examination were scored using the following formula: the number correct minus one-fifth the number of items omitted. The minimum passing score of 90 is approximately equivalent to a score of 119 on the AGCT.

[10]Requirements for the three major aircrew positions include: <u>Pilot</u> - ability to make quick and accurate observations and judgments, speed of reaction, complex motor skills, gross muscular coordination, ability to command, and confidence and aggressiveness; <u>Navigator</u> - superior general ability, understanding of abstract mathematical relationships, ability to make rapid and accurate mental calculations, ability to maintain spatial orientation with the use of instruments and maps, and some degree of muscular coordination; <u>Bombardier</u> - ability to concentrate, ability to make rapid mental calculations, ability to learn theory and operation of the bomb site, eye-hand coordination, finger dexterity, and motor steadiness.

the following ability or personal characteristic areas: (a) intelligence, judgment, and scholastic or educational achievement; (b) alertness, observation, and speed of perception; (c) temperament; and (d) psychomotor abilities. For the most part, we will focus on the research and results related to the cognitive abilities area (i.e., areas a and b above).

Tests designed to tap intelligence, judgment, and scholastic proficiency included measures of mathematics ability, numerical facility, reading ability, ability to interpret technical data, mechanical comprehension, and general knowledge. Measures related to alertness, observation and speed of perception included the ability to make rapid, accurate observations from information provided in maps, photographs, tables, and charts.

Throughout the war years, the tests comprising the classification battery were continually undergoing revision and modification. During this period, well over 200 tests were developed and psychometrically evaluated (Staff, Psychological Branch, 1943). In general, the Classification Battery contained 18 tests, 12 being paper-and-pencil measures of cognitive abilities and 6 measuring psychomotor skills. Detailed descriptions of the numerous experimental measures designed to tap these abilities may be found in Printed Classification Tests, Report 5 of the series published by the Aviation Psychology Program (Guilford & Lacey, 1947).

The Classification Battery was administered over a two-day period in which examinees completed the cognitive tests on the first day and the psychomotor tests on the second day. In addition, examinees were asked to rank order their preferences for the bombardier, pilot, and navigator positions. Classification tests were then scored and four aptitude or composite scores were computed (scores for the pilot, navigator, and bombardier positions and a total score). All scores were converted to stanine values.[11] Because the tests contained in the battery were constantly under revision, the passing stanine values used for selection and classification purposes were also revised or modified. Toward the end of the war, the passing or acceptable stanine value was set at six for pilots and bombardiers and at seven for navigators.

During the war, approximately 600,000 men completed the AAF Aviation Classification Battery. About 42 percent qualified for pilot training, 9 percent for navigator training, and 9 percent for bombardier training; eight percent were disqualified for physical reasons, and 17 percent were assigned to other aircrew positions such as radar observers, flight engineers, mechanics, and gunners. Thus, approximately 16 percent of those completing the classification battery were rejected on the basis of low aptitude or ability scores (Flanagan, 1947).

---

[11]The stanine ("standard nine") score, developed by the AAF Aviation Psychology Program research group, represents a standardized score. The aptitude score or stanine values possess a mean of five, a standard deviation of two, and a range from one to nine. These scores are designed to represent a normal distribution.

Stanine Score	1	2	3	4	5	6	7	8	9
Normal Curve Percentile	4	7	12	17	20	17	12	7	4

To classify examinees obtaining one or more passing aptitude scores, a board was given information about each person's score and position preference. Aircrew assignments were then made by matching a person's preference for one of the aircrew positions and his aptitude scores indicating the aircrew assignment for which he would be best suited.

As noted above, throughout this program literally hundreds of tests were developed, administered, and correlated with some measure of job performance; Classification Battery tests were constantly being revised or omitted and new ones added. Thus, it is difficult to summarize the validity data for all tests included in the battery. Instead, Table 11 contains validities for tests and composite aptitude scores obtained from the December 1943 Classification Battery. These data represent correlations between test scores and training outcome scores for each aircrew position. Multiple correlations between aptitude or composite test scores and training outcome scores are also presented for each position. Descriptions of the measures included in this table are provided in Appendix A.

According to the results presented in Table 11, the most effective cognitive measures for predicting success in pilot training are Reading Comprehension, Spatial Orientation, Dial and Table Reading, Mechanical Principles, Technical Vocabulary, and Instrument Comprehension. For bombardiers, whose performance is not predicted as well as performance in the other two jobs, the most effective cognitive measures include Reading Comprehension, Spatial Orientation, Dial and Table Reading, Numerical Operations, and Arithmetic Reasoning. For navigators, whose performance was 22predicted best, this list includes Reading Comprehension, Spatial Orientation, Dial and Table Reading, Arithmetic Reasoning, and Numerical Operations. Also, note that for the pilot and navigator positions, interests, background data, and attitudes predict success in training. Finally, across the three aircrew positions, several psychomotor ability tests effectively predicted training success.

A Predictive Validity Study. To evaluate the effectiveness of both the AAF Cadet Qualifying Examination and the Aviation Classification Battery, the Aviation Psychology research group obtained approval in mid-1943 to conduct a "pure" predictive validity study. In this study, an experimental group consisting of 1,305 applicants completed the qualifying exam and classification battery. Scores on these measures were not used to make accept/reject or classification decisions.

Instead, all applicants who passed the physical (N = 1,142) were accepted into preflight pilot training school regardless of their test scores. Later analysis of Cadet Qualifying Examination test scores revealed that 58 percent of the experimental group obtained passing scores while 42 percent would have been rejected on the basis of their scores (Flanagan, 1947).

Table 11

Aviation Classification Battery (December 1943) Subtest Validity Coefficients for Three Aircrew Jobs

Measure	Pilot		Bombardier		Navigator	
	r	N	r	N	r	N
Cognitive/Perceptual						
Reading Comprehension	.19	7,400	.12[a]	3,200	.32[b]	400
Spatial Orientation I	.20[c]	9,100	.12	3,200	.38[b]	700
Spatial Orientation II	.25[c]	9,100	.09	3,200	.33[b]	700
Dial and Table Reading	.19[c]	3,200	.19[a]	3,200	.53[b]	700
Mechanical Principles	.32[c]	8,100	.08	1,800	.13	300
Technical Vocabulary (Pilot)	.30[c]	13,700[d]	.04	3,200	.10	700
Technical Vocabulary (Navigator)	.09	13,700	.04	3,200	.22	700
Mathematics	.08	16,300	.10	3,200	.50	--[d]
Arithmetic Reasoning	.09	10,500	.12[a]	3,200	.45[b]	--[d]
Instrument Comprehension I	.15[c]	600	--	--	--	--
Instrument Comprehension II	.35[c]	600	--	--	--	--
Numerical Operations, Front	.01	9,100	.13	3,200	.26	1,500
Numerical Operations, Back	.02	9,100	.11	3,200	.28	1,500
Speed of Identification	.18	20,100	.09	3,200	.19	1,500
Psychomotor/Apparatus						
Rotary Pursuit	.21[c]	8,100	.14[a]	1,800	.10	700
Complex Coordination	.38[c]	24,100	.18[a]	3,200	.24	700
Finger Dexterity	.11	15,200	.16[a]	3,200	.20[b]	700
Discrimination Reaction Time	.22[c]	13,700	.16[a]	3,200	.36[b]	700
Two-Hand Coordination	.30[c]	12,500	.12	2,200	.26[b]	700
Rudder Control	.42	1,000	--	--	--	--
Biographical Data Pilot	.32[c]	7,000[d]	--	--	--	--
Biographical Data Navigator	--	--	--	--	.23[b]	300
Multiple R	.57		.29		.69	

Note: From The Classification Program (p. 99) by P. H. DuBois (Ed.) (1947), Washington, DC: Army Air Force Aviation Psychology Program Reports, No. 2.

[a]Subtests included in a computation of the Bombardier Aptitude Score.
[b]Subtests included in the computation of the Navigator Aptitude Score.
[c]Subtests included in the computation of the Pilot Aptitude Score.
[d]Estimated value from various forms of the measure or from several samples.

Members of the experimental group were followed from preflight training through primary, basic, and advanced pilot training. Of the total number accepted into the preflight program, 23 percent (N = 265) actually completed all pilot training programs and became certified pilots. Analyses of classification battery pilot stanine scores indicated that for applicants obtaining stanine scores of one, two, and three, only about 3 percent became certified pilots; for those obtaining stanine scores of four, five, and six, about 29 percent completed all pilot training programs, whereas for those obtaining scores of seven, eight, or nine, 60 percent became certified pilots.

For the entire experimental group, pilot stanine scores were correlated with graduation or elimination from advanced pilot training, yielding a validity coefficient of .65. Subtest scores for the complete classification battery, when combined to produce a maximally weighted linear sum, yielded a multiple correlation of .67 with the graduation/elimination criterion measure. Using the same criterion measure, the best weighted sum of all paper-and-pencil measures yielded a multiple correlation of .61 while the best weighted psychomotor test composite produced a multiple correlation of .57. A maximally weighted composite of Cadet Qualifying Examination subtest scores yielded a multiple correlation of .48. Finally, the pilot stanine and Cadet Qualifying Examination score, when combined, produced a multiple correlation of .65 (DuBois, 1947).

According to the results of this predictive validity study, the pilot stanine score derived from Aviation Classification Battery subtest scores effectively predicts success in the Army Air Force pilot training program. In addition, the pilot stanine appears to be more effective than the best weighted composite of paper-and-pencil measures and the best weighted composite of psychomotor measures. Further, the pilot stanine appears to work as well as the best weighted composite of all classification battery subtests (e.g., pilot stanine $\underline{r}$ = .65 versus Classification Battery $\underline{r}$ = .67).[12]

Finally, the Cadet Qualifying Examination appears to add little to the prediction of pilot training success when combined with the pilot stanine score (i.e., pilot stanine alone $\underline{r}$ = .65 vs. pilot stanine plus Qualifying Examination $\underline{r}$ = .66). DuBois contended that the advantage of using the Qualifying Exam to screen pilot applicants was not related to the unique variance it added to the predictor equation, but involved time and cost savings in administering a three-hour test versus a two-day battery of tests to eliminate potentially unsuccessful applicants.

To summarize the results for cognitive abilities from the Army Air Force Aviation Psychology Program, we have prepared a table that lists and defines all cognitive constructs identified as potentially important for success in

---

[12]DuBois (1947) notes that the multiple correlations computed for the entire classification battery subtests for paper-and-pencil measures only, and for psychomotor measures only, were not cross-validated. Thus, the amount of shrinkage occurring for each value is unknown. Less shrinkage, however, would be expected for the pilot stanine score because it does not involve a maximally weighted composite that capitalizes on chance.

aircrew performance (see Table 12.) Also included in this table are the target aircrew positions for which measures of each construct proved valid. This information represents a summary analysis of the data collected by Aviation Psychology Program researchers throughout the duration of the war (Guilford & Lacey, 1947).

Note that all cognitive constructs are linked to performance in either pilot or navigator positions or in both; fewer constructs are linked to performance in the bombardier position. Also note that, with the exception of reading comprehension, constructs useful for predicting training outcomes across the three aircrew positions relate to perceptual abilities (perceptual speed), or spatial abilities (visualization).

Motion Picture Testing. Although the validity of perceptual ability measures of aircrew performance was documented near the end of the war, the importance of these abilities became clear to aviation researchers very early in the design of the program. Therefore, a research unit specifically geared toward developing measures of perceptual abilities was established early in the Aviation Psychology Program. The goal of this unit, the Motion Picture Testing Program, was to develop measures related to assessing and evaluating visual cues and to present these measures in a more realistic fashion than was possible with paper-and-pencil measures. Motion picture films were developed for selection, training, and job proficiency testing purposes. The films were designed to correspond to, or more realistically represent, events that arise in aerial combat situations. We focus here on the measures designed for selection and classification purposes.

Results from job analyses of aircrew performance provided information about the perceptual abilities or functions that are required for success in these occupations. Members of the Motion Picture Testing Program used this information to identify several perceptual ability constructs that could not be measured adequately by traditional paper-and-pencil tests, but could be captured more effectively in motion picture tests. Eight perceptual ability constructs were identified: ability to judge motion and locomotion, ability to judge distance, ability to maintain orientation in space, ability to perceive slight movement, ability to perceive multiple stimuli, ability to perceive and integrate sequentially presented material, speed of perception, and comprehension of verbal and visual instructions (Gibson, 1947).

To develop measures of these constructs, researchers first determined the item types for inclusion in each test and then screened available film footage or planned for specific footage to be filmed. All tests were designed to provide instructions directly on the film. In general, the film tests contained several multiple-choice items to which subjects responded on machine-scorable answer sheets. Overall, 15 tests were developed. Because

Table 12

Cognitive Ability Constructs Assessed by Airman Classification Battery Subtests

Construct	Definition	Target Air Crew Positions[a]		
Verbal Ability	Viewed as a general intelligence or conceptual intelligence measure.	N		
Reading Comprehension	Ability to read and comprehend material related to pilot, bombardier, and navigator activities.	P	N	B
Mathematical Ability	Indicative of abstract intelligence, ability and achievement in advanced arithmetic, algebra, and trigonometry.		N	B
Number Facility	Measures simple arithmetic processes.		N	B
General Reasoning	Ability to accurately reason with words and numbers.		N	
Analogical Reasoning	Ability to reason with figures (non-verbal and non-numerical ability).	P		
Judgment	Ability to react immediately and appropriately to stimuli; ability to grasp the situation as a whole.	P		
Planning	Being fully prepared and fully briefed about a situation, knowledgeable of what to do in an emergency situation.	P	N	
Integration	Ability to construct an integrated impression; ability to keep all elements in a set operating effectively.	P	N	
Memory	Ability to absorb large quantities of material, meaningful or meaningless, in a short amount of time.	P	N	
Visual Memory	Ability to remember and to recognize material of a non-verbal pictorial nature.	P	N	
Symbolic Memory	Ability to remember meaningful material over a long term.	P	N	
Visualization	Ability to mentally manipulate visual images.	P	N	B
Mechanical Comprehension	Ability to succeed in pursuits involving operation and utilization of mechanical equipment.	P	N	
Perceptual Speed	Ability to rapidly and visually assess detail or to recognize similarities and differences.	P	N	B
Form Perception	Ability to reorganize disordered segments into a coherent whole.	P		
Size and Distance Estimation	Ability to accurately perceive size and distance of objects.	P	N	
Spatial	Ability to make discriminations as to direction of movement, and as to position of objects.	P	N	B
Orientation	Ability to determine one's bearings with respect to points of a compass and ability to maintain or establish location relative to landmarks in the environment.	P	N	
Set and Attention	Ability to concentrate or sustain mental effort; ability to resist distraction (divided attention) and ability to change mental set in approach to new problems.		N	

Note: Summarized from Printed Classification Tests by J. P. Guilford and G. I. Lacey (Eds.) (1947), Washington, DC: Army Air Force Aviation Psychology Research Program Reports, No. 5.

[a] P = Pilots, N = Navigators, B = Bombardiers.

much of the test development in this area was completed near the end of the war, only a few measures were actually administered and validated against a criterion of graduation versus elimination from elementary pilot training.[13] A description of the 15 measures and results from available pilot studies are presented in Table 13.

Results provided in Table 13 indicate that measures of ability to judge motion and locomotion, sequential perception, and comprehension of visual and vocal instructions yielded only low to moderate reliability estimates (range .34 to .68). Reliability estimates for measures of the ability to judge distance, perception of slight movement, multiple perception, and quickness of perception are higher (range .53 to .94). For the construct ability to maintain orientation in space, no data are available. Concerning available validity estimates, measures of multiple perception and quickness of perception appear to be most useful in predicting pilot training outcomes. Of particular interest are the measures of multiple perception--flexibility of attention and integration of attention. These two measures appear to have more general applicability in predicting success in occupations other than aircrew performance.

Although some motion picture measures appear to be potentially useful for selection purposes, little is known about their practical utility. Further, because most tests were not completed until the end of the war, none of these measures were actually incorporated into the Aircrew Classification Battery. Therefore, it is unclear whether these measures would add unique variance to the prediction of training success for any of the aircrew positions. In sum, research conducted by the Staff at the Personnel Research Section and by the Aviation Psychology Program marked great strides in selection research. First, measures were developed to assist with the initial selection of military personnel. Second, numerous aptitude and knowledge tests were developed to aid in classifying personnel into literally thousands of military occupations; unique testing procedures such as motion picture tests were developed and their effectiveness for selection purposes documented. Finally, test development procedures used in these research programs remain virtually unchanged from those recommended today in test development and validation research. Perhaps the contribution made by these research programs can best be summarized by the following:

> It has been generally recognized that to the U.S. Army belongs the credit for developing personnel methods which have since been widely copied by

---

[13]A series of studies was conducted to determine the influence of viewing distance and viewing angle on test performance. For three measures (flexibility of attention, integration of attention, and minimal movement) viewing distance was significantly related to test performance. Viewing angle, however, was not related. Additional studies were conducted to assess the effect of room illumination on test performance. Results indicated that extreme high and extreme low illumination levels did not affect performance, although lower illumination levels appeared optimal for this type of test administration.

Table 13

Motion Picture Film Measures: Test Descriptions and Results

Construct/Measure	Description
Ability to Judge Motion And Locomotion	
Estimate of Velocity	Capacity to estimate and visualize speed of an object moving at right angles. $r_{xx'}$ = .50 - .65 Internal Consistency r_{xy}[a] = .00 - .05 (N range 250-750; median r = .03)
Identification of Velocity	Ability to discriminate visual velocities in a Velocity relatively "pure" form. $r_{xx'}$ = .44 - .61 Internal Consistency r_{xy}[a] = .07 - .16 (N range 250-767; median r = .12)
Estimation of Relative Velocities	Requires complex judgment to ascertain the Velocities relation between two objects. $r_{xx'}$ = .34 - .67 Internal Consistency r_{xy}[a] = .03 - .21 (N range 250-1047; median r = .14)
Landing Judgment	Ability to learn certain spatial discriminations believed required for successfully landing a plane. $r_{xx'}$ = .34 test-retest No validity data available.
Ability to Judge Distance	
Distance Estimation	Ability to make spatial discriminations based on perception of distance. $r_{xx'}$ = .57 to .79 Internal Consistency No validity data available.
Ability to Maintain Orientation in Space	
Flying Orientation	Ability to maintain directional orientation when flying and ability to visualize a flight path. No data collected on this measure.
Landing Orientation	Ability to discriminate, learn, and remember the features of the ground that serve as cues for spatial orientation in the traffic pattern. No data collected on this measure.
Perception of Slight Movement	
Minimal Movement	Ability to detect barely visible movement of an object and to determine the direction of this movement. $r_{xx'}$ = .69 - .77 Internal Consistency No validity data available.
Drift Detection	Ability to detect drift of a moving spot to one side or the other of the main direction in which it moves. $r_{xx'}$ = .59 - .62 Internal Consistency No validity data available.

(Continued)

Table 13 (Continued)

Motion Picture Film Measures: Test Descriptions and Results

Construct/Measure	Description
Multiple Perception	
Flexibility of Attention	Ability of an aircrew candidate to distribute attention over a wide range of stimuli.
	$r_{xx'}$ = .53 - .94 Internal Consistency
	r_{xy}^{a} = .05 - .26 (N range 219-1097; median r = .15)
Integration of Attention	Ability to distribute attention over a complete field of events and to treat this field as an interconnected whole.
	$r_{xx'}$ = .71 - .88 Internal Consistency
	r_{xy}^{a} = .07 - .15 (N range 296-1097; median r = .09)
Sequential Perception	
Successive Perception	Ability to integrate successive partial impressions into a single visual scheme or pattern.
	$r_{xx'}$ = .34 - .55 Internal Consistency
	No validity data available.
Successive Perception Test II	Ability to form an integrated total impression of a visual experience which has been in perceived in successive stages or parts.
	$r_{xx'}$ = .48 - .68 Internal Consistency
	No validity data available.
Quickness of Perception	
Plane Formation	Ability to apprehend a visual pattern within a brief exposure period and reproduce it accurately.
	$r_{xx'}$ = .82 Internal Consistency
	r_{xy}^{a} = .12 - .22 (N range 250-956; median r = .16)
Comprehension of Visual and Vocal Instructions	
Motion Picture Comprehension	Ability to comprehend and remember material which is presented in motion picture form with visual demonstrations and diagrams accompanied by an explanatory narrative.
	$r_{xx'}$ = .63 Internal Consistency
	No validity data available.

Note: From Motion Picture Testing and Research by J. J. Gibson (1947), Washington, DC: Army Air Force Aviation Psychology Program Reports, No. 7.

[a] The criterion measure consisted of graduation versus elimination from elementary pilot training.

the armies of other nations and which have had an important effect upon the progress of comparable civilian work (Staff, Personnel Research Section, 1943a, pp. 129-130).

Current Military Selection and Classification Battery

Post World War II. Research conducted on tests developed during World War II provided the necessary data to develop and implement more systematic screening systems after the war. For example, classification tests constructed and validated during the war--such as the AGCT, Mechanical Aptitude, Clerical Speed, Radio Code Learning, and Automotive Information, along with others--were combined to form the Army Classification Battery (ACB). This battery, containing ten subtests, became operational in 1949. Validity data collected on each of the subtests during the war were used to identify different combinations of subtests to predict success in different occupations. Thus, scores on ACB subtests were used to compute ten aptitude area scores. These aptitude area composites, consisting of two or three subtest scores, were used to classify recruits into one of ten broad occupational areas.

Further, data obtained from a shortened version of the AGCT were used to develop a selection test for all recruits entering the Army. This measure, R-1, became operational in 1946. Shortly thereafter, in 1948, the passage of the Selective Service Act generated a need for uniformity in mental testing procedures for all services (Uhlaner, 1952). The Office of the Secretary of Defense authorized a committee of Army, Air Force, and Navy personnel to develop uniform screening tests and scoring systems for all inductees and enlistees in the Armed Force. Efforts by this joint committee resulted in the Armed Force Qualification Test (AFQT). The first version of this selection test, AFQT 1 and 2, contained items similar to those in the AGCT (i.e., verbal, arithmetic reasoning, and spatial); later versions contained an additional measure, Tool Usage.

The AFQT became operational as a selection device for all branches of the Armed Force in 1950. As for the AGCT, scores on this measure were converted to percentile values and grouped into five mental ability categories. Raw scores on the AFQT were normed against a World War II reference population consisting of 12 million officers and enlisted personnel. Thus, the AFQT mental categories and percentile scores yield a distribution similar to that reported for the AGCT (see Table 14). The AFQT, in several revisions, was used by all Armed Force branches until 1972 when it was discontinued; each Service at this point used its own selection test battery.

Armed Services Vocational Aptitude Battery (ASVAB). Further modifications of the Armed Force selection program began in 1966 when the Assistant Secretary of Defense, Manpower, and Reserve Affairs established a Joint Services committee. This committee was charged with developing and standardizing a single high school aptitude battery to meet the needs of all branches of the Armed Force (Vitola, Mullins, & Croll, 1973). Resulting from this joint effort was the Armed Services Vocational Aptitude Battery (ASVAB) containing subtests constructed from items included in Army, Navy, and Air Force classification tests. In September 1968, the ASVAB became operational

Table 14

Armed Force Qualification Test (AFQT) Category Ranges and World War II Reference Population Distribution

AFQT Category	Percentile Score Range	World War II Reference Population Distribution[a] (Percent)
I	93-100	7
II	65-92	28
III	31-64	34
IV	10-30	21
V	1-9	10
		100

Note: From *Screening for Service: Aptitude and Education Criteria for Military Entry*, by M. J. Eitelberg, J. H. Laurence, B. K. Waters, & L. S. Perelman (1984), Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Installation and Logistics)

[a]The reference population approximates the aptitude score distribution of males on active duty (including 12 million officers and enlisted personnel) as of 31 December 1984.

in the military high school testing program. In 1976, the ASVAB was implemented as the single Department of Defense enlistment test.

The ASVAB, like the earlier selection and classification batteries, undergoes revision on a continuing basis. As of 1989, the battery contained 10 subtests (ASVAB 15, 16, and 17), which are listed and described in Table 15. Four measures--Word Knowledge, Paragraph Comprehension, Mathematical Knowledge, and Arithmetic Reasoning--are used to compute the current Armed Force Qualification Test (AFQT) score, the score used to determine enlistment eligibility.

Procedures used to norm the original AFQT in 1950 were also used to norm AFQT scores. Thus, percentile scores derived from this measure may be interpreted similarly to those derived for the 1950 AFQT and the AGCT (see Table 14)[14]. Enlistment eligibility is determined by AFQT percentile score or mental ability category (i.e., I-V), along with information about education achievement (i.e., high school graduate versus non-high school graduate), and results from a physical examination and morals screening.

Occupational classification of an enlistee is determined by the various Services according to scores on Aptitude Area composites, which are a combination of scores obtained from three to five ASVAB subtests. Nine Aptitude Area scores used by the Army to represent nine broad occupational groups are computed for each enlistee. The ASVAB subtests used to derive each aptitude area score are identified in Table 16. Aptitude Area scores are then used to assess enlistees' qualifications for assignment into each of the broad occupational groups and into a particular Military Occupational Specialty (MOS). Examples of MOS within each aptitude area are provided in Table 16.

<u>Summary</u>

Military research during World War II provided a wealth of information for designing selection and classification systems. Staff of the Personnel Research Section, responsible for developing selection and classification measures for the U.S. Army, expanded upon the previous group-administered test, the Army Alpha. Numerous tests and batteries were developed for non-English speaking applicants, applicants failing to meet minimum educational requirements, and personnel with special skills or with officer potential. By the end of the war, well over 12 million enlisted personnel and officers had been screened and classified using one or more of these measures (Eitelberg, et al., 1984).

Researchers for the Army Air Force constructed hundreds of tests to screen and classify applicants into aircrew positions. The methodology that resulted appears to have been a model for the current military selection and classification battery. For example, aircrew applicants were first screened

---

[14]Test score data from the <u>Profile of American Youth FY 80</u> (Department of Defense, 1982) were used to renorm AFQT scores using a nationally representative sample of youth. Unlike the 1944 reference population, this sample includes approximately equal numbers of males and females.

Table 15

Descriptions of Subtests Included in the Armed Services Vocational Aptitude Battery (ASVAB)

ASVAB Subtest	Cognitive Ability
Word Knowledge (WK)	Ability to understand the meaning of words.
Paragraph Comprehension (PC)	Ability to read and understand written material.
Numerical Operations (NO)	Ability to quickly and accurately perform simple arithmetic operations.
Arithmetic Reasoning (AR)	Ability to solve mathematical word problems.
Coding Speed (CS)	Ability to perceive visual information quickly and accurately and to perform simple processing with it.
Mathematics Knowledge (MK)	Ability to correctly use algebraic formulae to solve problems.
General Science (GS)	Knowledge of science information acquired in high school courses.
Mechanical Comprehension (MC)	Ability to comprehend and reason with mechanical terms.
Electronics Information (EI)	Knowledge and understanding of electricity, radio, and electronics.
Auto and Shop Information (A/S)	Knowledge and understanding of automobiles, tools, and shop practices.

Table 16

ASVAB Subtests Used to Compute Aptitude Area Scores

Aptitude Areas	ASVAB Subtests									
	WK	PC	NO	AR	CS	MK	GS	MC	EI	A/S
Combat (e.g., Infantryman - 11B)	x	x			x		x			
Field Artillery (e.g., Cannon Crewman - 13B)				x	x	x		x		
Electronics Repair (e.g., Tow and Dragon Repairer - 27E)				x		x	x	x		
Operators and Food Handlers (e.g., Motor Transport Operators - 64C)	x	x	x					x		x
Surveillance and Communication (e.g., Radio/Teletype Operator - 31C)	x	x		x				x		x
Mechanical Maintenance (e.g., Light Vehicle Repair - 63B)			x					x	x	x
General Maintenance (e.g., Ammunitions Specialist - 55B)						x	x		x	x
Clerical (e.g., Administrative Specialist - 71L)	x	x	x		x					
Skilled Technical (e.g., Medical Specialist - 91B)	x	x				x	x	x		
General Technical[b]	x	x		x						

[a]Subtest Abbreviations:

WK = Word Knowledge
PC = Paragraph Comprehension
NO = Numerical Operations
AR = Arithmetic Reasoning
CS = Coding Speed
MK = Mathematical Knowledge
GS = General Science
MC = Mechanical Comprehension
EI = Electronics Information
A/S = Auto Shop Information

[b]This composite is not used to make classification decisions. Instead it is used to determine reenlistment qualifications or special educational needs.

using the Cadet Qualifying Exam; today, the Armed Force Qualification Test is used to screen all military applicants. Within the Army Air Force, scores on the Classification Battery were used to assign qualified applicants to various aircrew positions; today, composite scores on ASVAB subtests are used to match individual abilities with job requirements.

Research conducted by the Army Air Force staff led to the expansion of the cognitive ability domain. Tests developed by this group were initially derived from Thurstone's seven primary cognitive abilities. Subsequent research, however, indicated that many more cognitive ability constructs could be identified. Following the war, Guilford continued to explore the cognitive ability domain, later proposing the existence of more than 120 abilities.

Finally, Army Air Force research staff demonstrated that perceptual ability tests administered via motion pictures could also be used to expand the cognitive ability domain. Although only a few measures were validated before the end of the war, the available validity data suggest potential for these measures in a selection setting.

WORLD WAR II: CHANGES IN EMPLOYMENT PRACTICES AND EMPLOYMENT OPPORTUNITIES

Another event occurring during World War II that had significance for hundreds of thousands of individuals was the dramatic need for personnel to staff war production plants. This need, in conjunction with the great numbers of men who volunteered or were drafted for military service, created job opportunities for nearly anyone who wished to work. Thus, women and minorities were hired to fill what had traditionally been white male occupations. Similar opportunities were available for women and minorities in the military sector, although the increase in the numbers of women and minorities in non-traditional military occupations appears to have been less dramatic than the increase in the private sector. As a result of these opportunities, women and minorities experienced changes in occupational interests and in employment expectations. In this subsection we examine these changes in both the private sector, or the homefront, and the military sector, and discuss their implications for future employment practices.

The Homefront

For women, employment practices during World War II, unlike those during the first world war, offered great numbers as well as variety in employment opportunities. Although women were encouraged to participate in the war effort during World War I, jobs available to them were restricted to traditionally female occupations, such as secretary and clerk. Further, employers limited hiring to young, unmarried women. Thus, the increase in employment for women during the first world war did not produce dramatic changes in hiring practices. Following the war, women were expected to leave their jobs voluntarily. If they did not, they were terminated without question, to ensure that jobs were available for men returning from the war (Anderson, 1951).

In 1941, the Lend-Lease Act as well as the declaration of war produced an ever-increasing need for workers in war production plants. Employers no longer hired only young, unmarried females. Instead, all women, young and old, unmarried and married, were encouraged to work in these plants; industries mounted extensive campaign efforts to recruit them. Women were no longer limited to traditionally female occupations. Instead, they were hired to serve as blue-collar workers, such as precision tool makers, overhead crane operators, lumberjacks, drill press operators, stevedores, and switch operators (Anderson, 1951). Women were also hired to serve in white-collar occupations traditionally reserved for males. They began working as journalists, radio personalities, symphony orchestra members, and stock brokers. Prior to the war, women comprised 25 percent of the work force. By 1944, the peak year of female wartime employment, they constituted 36 percent of the work force (Harris, Mitchell, & Schechter, 1984).

Although women were hired in great numbers and were successful in a wide variety of occupations, common employment practices relating to women prevailed. For example, women were often overlooked for promotions and were discouraged from taking exams that would lead to job advancements (Anderson, 1951). Employers viewed women's contribution in the work effort as less valuable than men's contributions and, therefore, offered lower wages to women for the same work. Ironically, some unions pressed for equal pay for women; union leaders feared that women would be hired instead of men, because employers paid women less (Harris et al., 1984).

At the outset of World War II, blacks were still often barred from applying for jobs traditionally held by white males. In 1941, the president of the Brotherhood of Sleeping Car Porters and other black leaders called for a march on Washington to protest the lack of job opportunities for blacks in defense plants. The march was cancelled after President Roosevelt issued Executive Order 8802, banning discrimination in defense industries and government based on "race, creed, color or national origin" (Harris, et al., 1984).

The Fair Employment Practices Committee (FEPC) was established to enforce the ban. Members of the Committee were tasked with conducting hearings to assess and evaluate defense contractors' employment practices. A member of that committee, Earl B. Dickerson, cited some examples of the employment conditions for blacks during that period: (a) a subsidiary of a large automobile manufacturing plant reported having one black in their employment, and (b) a large defense contractor located in California with 20,000 employees reported having no blacks on their rosters (Terkel, 1984).

Black leaders continued to work along with members of the FEPC during the war to ensure that jobs opened up to blacks. Alexander Allen, industrial relations director of the Baltimore Urban League, described the situation for blacks in Baltimore during the war. "In 1942 the number of blacks in manufacturing industry was nine thousand. By 1944 they had increased to thirty-six thousand." This represents an increase from six percent to 15 percent of blacks in the work force (Harris et al., 1984).

Unfortunately, the end of the war triggered a sharp decrease in job opportunities for blacks and women. For blacks, the "last hired, first fired" rule applied. For example, in Baltimore the number of blacks employed in the manufacturing industry following VJ Day decreased to 12,000, or 12.5 percent of the total work force.

Following the war, women were encouraged to resign from their jobs, thereby providing men returning from the war with jobs. Although considerable pressure was applied to women to surrender their jobs, a 1944 Labor Department study reported that 80 percent of the women interviewed wanted to continue working in some kind of job after the war (Harris et al., 1984).

Conditions in the Military Sector

As noted previously, women volunteered for military duty during this time. In fact, a special selection battery, the Women's Classification Test, was developed to screen women entering the Army. Researchers involved in the Aviation Psychology Program reported that, when women volunteered for duty in the Army Air Force, the Aviation Classification Battery was used to select and classify women into pilot positions. Detailed information concerning male-female differences in test battery scores is not provided. The authors concluded, however, that, although differences appeared on some measures, especially those related to mechanical comprehension, the Aviation Classification Battery tests appeared to effectively predict aircrew performance equally well for men and women (Flanagan, 1947). Very little additional information describing women's roles and activities in the military during this period is available.

Conditions for blacks in the military appear to have been somewhat bleaker than those for women, especially during the early years of the war. During this period, the Armed Services were segregated; blacks were prohibited from using white recreation and PX facilities and often had no such facilities available for their own use. (In 1948, President Truman ordered the desegregation of the Armed Forces and a ban on discrimination in federal jobs.)

Concerning job assignments, black GIs were often restricted to labor battalions, assigned menial duties, and excluded from officer ranks (Terkel, 1984). As the war progressed, changes in military policy resulted in classifying a small number of blacks into more challenging occupations. For example, all-black tanker crews were established and used in the European front; blacks were included in the Army Air Force as pilots, although the total number was minute compared to the number of white males serving as pilots[15].

---

[15]Of the 600,000 men who completed the AAF Aviation Classification Battery, 42 percent, or 252,000, qualified as pilots. A total of 996 blacks served as pilots in the Army Air Force (Guilford & Lacey, 1947; Terkel, 1984).

For both blacks and women, World War II paved the way for opportunities to work in a wide variety of well-paying jobs. Hence, their expectations about occupational opportunities and wages or salary changed from prewar times. Although the end of the war signaled a return to the earlier status quo, female and black group leaders continued to work for occupational equality with white males.

In the area of selection and placement, changes in the composition of the work force ultimately led to concerns about test usage. For example, do selection tests discriminate against females and blacks and thereby prohibit them from entering traditionally white-male occupations? Do selection test scores provide different information for different subgroups? Should selection tests validated on a sample of white males be used to evaluate the potential of females and blacks for success on a job? Should women and blacks be considered along with males when making promotion decisions? These and other questions started initially as social concerns but later became legal issues as the government became more active in protecting the employment rights of blacks, other minorities, and women.

Summary

The final part of this section focused on changes in the work force that prevailed during the war years. As noted, jobs normally available only to white males, became accessible to females and blacks. Changing the composition of the work force did not come easily for employers and employees; for example, government intervention was required in some cases to ensure jobs for blacks, other minorities, and females. On the other hand, many employers conducted extensive recruitment campaigns to attract non-traditional employees.

The end of the war resulted in a return to the earlier status quo in the work place; women were encouraged to quit to ensure jobs for males returning from the war and "last hired/first fired" policies resulted in job losses for blacks and other minorities. Experiences during the war, however, spawned numerous questions about hiring policies and the procedures used to determine occupational and promotional eligibility.

SECTION SUMMARY AND CONCLUSIONS

The focus of this section has been on conserving human talent by matching relevant individual characteristics with job requirements to ensure that human talents are fully utilized in the work setting and that no one is underutilized (or overtaxed). Research conducted at the Employment Stabilization Research Institute (ESRI) demonstrated that employment potential (or person-job matches) may be determined by assessing a wide variety of personal characteristics, including educational status, intelligence, clerical aptitude, mechanical aptitude, manual dexterity, physical strength, vocational interests, temperaments, trade skills, and sensory acuity. Information on such variables was used to provide vocational guidance and counseling at ESRI.

During World War II, researchers developed procedures for isolating critical job requirements and linking these with cognitive abilities, psychomotor skills, or other personal characteristics. Researchers experimented with a variety of cognitive ability tests and demonstrated that many of these measures were successful in predicting job performance. Test development procedures as well as many of the tests can be directly tied to the current military selection and classification battery.

Measures contained in the current battery, the ASVAB, tap a variety of cognitive abilities and technical knowledge. Results from a study designed to isolate the underlying factors in this battery indicate that the 10 subtests tap verbal ability, speeded performance, quantitative ability, and technical knowledge (Kass, Mitchell, Grafton, & Wing, 1982). From our cognitive taxonomy, described in Section II (see Table 7, p. 46), it is clear that measures of other cognitive ability constructs--for example, measures tapping memory, spatial abilities, perception, and fluency--might be added to the screening battery without introducing overlapping or redundant measures.

A review of the cognitive ability measures used to predict aircrew performance during World War II suggests that other constructs could be added to the cognitive ability taxonomy. For example, results from the Aviation Psychology Program indicated that measures of spatial orientation and perceptual abilities assessed via motion pictures were useful in predicting aircrew performance. Measures such as these may succeed in adding unique variance to the prediction of performance in numerous Army military occupational specialties (e.g., armor crewman, infantryman, MANPADS, and cannon crewman).

A final comment in this section concerns changes in employment practices and employment opportunities available to females and blacks during World War II. During this period, testing for selection and classification purposes increased dramatically, but subgroup differences do not appear to have been the focus of research at that time. Concerns about possible discrimination in testing actually surfaced much earlier in the 20th century; the relation between scores on selection measures and employment decisions involving these subgroups did not become a target of formal study until the early 1960s. In the next section, we examine issues and data related to subgroup differences in cognitive ability scores.

SECTION IV

CONSERVATION OF HUMAN TALENT: PSYCHOMETRIC AND SOCIAL ISSUES IN COGNITIVE ABILITY MEASUREMENT

INTRODUCTION

As noted in the preceding section, the onset of World War II generated a need for greatly increasing the number of women in the work world. The demand for women to perform in traditionally male-oriented occupations led to questions about previous hiring practices that resulted in a greater variety of job opportunities for males than for females. In other words, assumptions about the distinction between "men's work" and "women's work" no longer appeared appropriate because the two were much less clearly defined.

Another equally important issue arising out of World War II involved employment opportunities for blacks. As described in the previous section, during the war, blacks were hired for jobs typically reserved for white males. Like women, blacks demonstrated that they could indeed perform effectively in jobs from which they had been restricted during prewar times.

Thus, questions related to ability differences between various subgroups were relevant to selection decisions. Although subgroup differences on mental ability tests had been a subject of study since the beginning of the testing movement, most early studies concentrated on differences between males and females and between blacks and whites. More recent studies have been undertaken to examine test performance differences in groups defined by racial and ethnic heritage.

As mental ability testing became more sophisticated and more widely used in educational and occupational settings, selection policies and their effects on racial, ethnic, and gender subgroup opportunities fell under closer scrutiny. Not until the mid-1960s, however, did the meaning and interpretation for subgroup differences in test performance become a necessary and legally required consideration for establishing educational and occupational selection standards.

In this section, we examine the evidence related to subgroup differences in cognitive ability test performance. This involves comparing mean test scores on cognitive ability measures for males and females and for different racial and ethnic subgroups. Also included is a discussion about the meaning of subgroup differences with respect to selection decisions. Next, we describe social and psychometric issues involved in using cognitive ability tests to make selection decisions. This includes a description of potential bias arising from test construction procedures (content bias) and statistical interpretation of test scores (differential validity and differential prediction). Data collected to support or refute both types of bias are summarized and discussed, and procedures to ensure fair use of test scores are defined. Finally, Federal regulations enacted to ensure equal employment opportunity for all subgroups are described. In addition, we describe Federal guidelines designed to aid in developing and implementing tests for selection and classification purposes and summarize key court decisions providing judicial interpretation of the Federal guidelines.

GROUP DIFFERENCES IN COGNITIVE ABILITY TEST PERFORMANCE: MALE AND FEMALE DIFFERENCES

Researchers have long recognized that males and females differ in many ways beyond the more obvious anatomical and physiological factors. One way to help clarify and characterize these differences has been to compare males' and females' mean test scores on measures of general intelligence and of specific cognitive abilities. Results from these comparisons indicate that, in general, males obtain higher scores on measures of some cognitive ability constructs while females outscore males on other cognitive ability measures.

Several theories have been postulated to explain the source of these differences. These include environmental factors involving early socialization that emphasizes different roles, activities, and pursuits for males and females (Sherman, 1967, 1974), genetic and/or hormonal differences that influence brain structure and brain organization (O'Connor, 1943; Resnick, 1982), and a combination of environmental and genetic factors. Whatever the reason for these differences, it is important to identify the cognitive ability constructs on which significant differences appear between males and females in order to assess how these differences influence selection decisions.

A substantial amount of literature reporting differences between males and females from infancy to adulthood is available. For example, from infancy to early childhood, males and females differ very little on measures of general intelligence. When differences do appear, females often score higher, in general, than males, but the difference is very small (Willerman, 1979). It is at this time, however, that females begin to excel in verbal fluency; they tend to begin talking earlier and develop a greater vocabulary than males of the same age (Maccoby & Jacklin, 1974). Recent evidence, however, suggests that when verbal fluency measures are administered to children in this age group, differences between males' and females' mean scores are mixed (Willerman, 1979).

Later, from childhood through adolescence, male-female differences begin to emerge on specific cognitive ability measures. For example, by age 8, males on the average obtain higher scores than females on measures of spatial ability and, by age 12, males outperform females on quantitative ability measures.

Although numerous differences between males and females in early and late childhood could be cited, the most relevant population is persons in late adolescence or young adulthood, the age group that is the target population potentially available to the Army for recruiting and enlistment. Thus, for purposes of this study, we examine male-female differences on measures of general intelligence and on specific cognitive ability measures for samples at high school and college age levels (i.e., 16 to 23 years of age). Before examining male and female test score differences, attention will be given to methodological problems that arise when interpreting subgroup differences in mean test scores.

Methodological Issues

Anastasi (1937, 1976) noted that when the mean test scores of two groups are being compared, several factors may influence observed differences. Mean differences may appear between two groups for reasons unrelated to the cognitive ability construct measured. Specifically, comparisons between males' and females' mean test scores may indicate that the two groups differ on cognitive ability measures because of (a) socialization factors, (b) selective factors, and (c) sample size effects.

Socialization Factors. Socialization factors include parenting differences in early childhood as well as differences in educational pursuits in later childhood and adolescence. For example, in early childhood males and females are traditionally encouraged by parents to engage in different types of activities. Females are often more sheltered and taught to be neater and quieter than boys (Anastasi, 1937). Females commonly are taught to nurture while males are encouraged to be more curious and self-reliant (Anastasi & Foley, 1949). According to this reasoning, play activities for females are more sedate than male play activities.

According to Anastasi (1937, 1976), socialization factors may influence performance on cognitive ability measures because of differential exposure to relevant environmental conditions. Below we have generated an example of how socialization factors may contribute to sex differences for the construct mechanical aptitude.

> In early childhood, males are encouraged to be more active and more curious while females are encouraged to be obedient and quiet. Thus, males have more opportunity to tinker with toys, to investigate how things work and to take things apart and put them back together. While in school males receive additional exposure to mechanical principles and properties in shop and electronics classes whereas females seldom enroll in these types of courses. Thus, in the area of mechanical aptitude, males have greater opportunity to work with and become familiar with principles governing mechanical operations. Because of this additional exposure, training and practice, males, on the average, score higher than females on measures of mechanical aptitude.

Educational curriculum differences may also produce differences between males' and females' scores on cognitive ability measures. For example, females have in the past been encouraged to focus less on science and mathematics and more on literature, art, and other "genteel" subjects (Anastasi, 1937).

Recent research in the area of mathematics indicates that parents and teachers can play a role in influencing a child's expectations of success in mathematics and perceptions of the value of mathematical study, and the likelihood that a child will enroll in higher level courses (Fennema &

Sherman, 1977; Haven, 1971; Parsons et al., 1983). For example, fathers have been found to emphasize different areas of study for male versus female children; fathers of sons report that advanced mathematics is important, whereas fathers of daughters report that verbal skills are more important (Parsons, Adler, & Kaczala, 1982). Interestingly, in this same study, mothers did not report emphasizing different areas for sons versus daughters. Ernest (1976) reported that after sixth grade, fathers were more likely to help children complete mathematics homework, even though mothers were more likely to help children with their homework, in general. Fox (1977, 1982) also reported that differences between males and females may be attributed to the lack of female role models in the field of mathematics; most advanced courses are taught by men. Unfortunately, even with a large body of research designed to locate environmental factors related to mathematical ability, most researchers would agree that it is still unclear if parental and teacher expectations and encouragement profoundly influence children's attitudes and achievement in mathematics (Benbow, 1988).

In another example, recent attempts have been made to equalize educational curriculum for males and females, especially in the areas of science. Evidence from one midwestern state suggests that females and males enroll in the same or very similar courses up to 11th or 12th grade. At this point, males continue to take science courses at more advanced levels while most females fail to enroll or drop out of these courses (Clark, 1983).

Selective Attrition. According to Anastasi (1976), selective elimination from high school occurs more frequently for lower ability students, and more males than females elect to drop out of high school before graduation. Thus, the test score distribution for a particular cognitive ability test administered to 11th- or 12th-grade students may not be representative of the true population of adolescents because lower ability students are missing. Further, the range of test score for males would be truncated, resulting in a negatively skewed distribution for males. Because of these hypothesized missing data points, then, males may obtain a higher mean score than females. In reality, the true mean score for the population of 16- to 18-year-old males may be equal to, or even lower than, the true mean score for females in the same age group. Hence, greater selective elimination for males than females may result in significant differences in mean scores that do not reflect real differences in the population.

Sample Size Effects. For a given population or group, mean test score performance varies from sample to sample. This variation or sampling error is greater for smaller than for larger samples; with small sample sizes, an observed mean difference between two groups varies or appears unreliable. To reduce sampling error and to ensure greater reliability of mean score differences between males and females, comparisons should be made using sufficiently large, representative samples. More reliable conclusions about observed sex differences can be drawn when results from different studies are accumulated and examined as a whole.

Another consideration involving sample size is its influence on the statistical significance of mean differences; with large sample sizes, very small differences between means may be statistically significant. As

Anastasi (1937) noted, a significant difference as well as a perfectly reliable difference does not preclude a large amount of overlap between two distributions. When a test is being used to make selection decisions, the amount of overlap between the distributions for males and females indicates whether disproportionate selection occurs. In other words, if males consistently obtain lower scores than females on a particular test and a small amount of overlap exists between the two distributions, then females will be selected more frequently than males. Thus, the selection ratio of females to males will be high. An index of overlap provides information about the similarity of two distributions and about the meaning that a significant difference between male and female mean test scores has regarding selection decisions.[16]

In a hypothetical example of the effects of large sample sizes, two cognitive ability measures have been administered to a sample of males ($N = 400$) and females ($N = 400$). Both measures have been standardized to yield a mean of 100 and a standard deviation of 16. In this example, females score significantly higher than males in Test A (110 vs. 90, $p < .001$) and on Test B (102 vs. 99, $p < .01$). The effect size, or _d_, however, for Test A ($d = 1.25$) indicates that only 11 percent of the males score at

---

[16]The effect size (d) or index of overlap is computed by the following:

$$d = \frac{\text{Mean 1 - Mean 2}}{\text{SD}}$$

Where d = effect size; Mean 1 = mean of the higher scoring group; Mean 2 = mean of the lower scoring group; SD = pooled estimate of the standard deviation.

The effect size (d) may be interpreted by using one of two procedures. In the first, d or effect size is used to derive Tilton's overlap statistic _O_ from tabled values. These values range from 0 (indicating no overlap) to 100 percent (indicating total overlap). This value is interpreted as the percentage of scores obtained by one group that may be matched by scores in another group (Dunnette, 1966).

The second procedure involves locating d (z) on a cumulative normal probability table and subtracting the obtained value from 1.00. The resulting value indicates the percentage of individuals in the lower scoring group reaching or exceeding the mean of the higher scoring group. This value ranges from zero (indicating that none of the lower group members reach or exceed the mean of the higher group) to 50 percent (indicating complete overlap or 50 percent of the lower scoring group reaching or exceeding the mean of the higher group) (Sevy, 1982). Although both procedures provide similar information about the two distributions, values obtained from the second procedure more clearly indicate how mean differences between two groups affect the selection ratio of one group over another. We use the second procedure throughout the remainder of this section to interpret effect size of subgroup mean differences.

or above the female mean. For Test B the effect size (d = 0.19) indicates that 43 percent of the males score at or above the female mean. Hence, greater overlap exists between males' and females' test score distributions on Test B than on Test A.

Results for both tests are depicted graphically in Figure 2. Note that disproportionate selection of females over males would be greater when using Test A than when using Test B. For example, if the cutting score is set at the mean for the total group, the ratio of females to males selected using Test A is greater than 2.34; using Test B, the selection ratio of males to females is 1.17 (Sevy, 1988). Thus, an index of overlap provides information about the effects of using a test to make selection decisions, beyond the information provided in a significance test for mean differences.

In sum, when comparing mean test scores for males and females, socialization and selective factors may influence obtained differences due to greater opportunity for one group to learn or practice tasks or due to different rates of selective attrition from the target population. These factors are difficult, if not impossible, to control in research. The third factor, sample size effects, is more easily controlled. As noted above, studies designed to examine male and female differences on cognitive ability or other types of measures should include fairly large, representative sample sizes (e.g., N = 150 or more for each subgroup). Further, when significant mean differences appear between males and females, computing effect size and an index of overlap provides information about whether disproportionate selection will occur when the test is used. Finally, pooling results obtained from several studies contributes to the reliability of conclusions about mean differences.

Mean Score Differences Between Males and Females: General Intelligence

In the area of general intelligence, mean test score differences between males and females appear but they are generally small and of little practical significance. For example, Yerkes (1921) administered the Army Alpha to male and female students attending normal school (two-year teachers' college) and college. The median Alpha score for males was higher than for females in both samples, but the differences were very small (males' mean score--115 for normal school, and 130 for college level; females' mean score--111 for normal school, and 127 for college level). Overall, Yerkes concluded that the differences between males and females on measures of general ability may be regarded as of little consequence.

Group-administered intelligence tests developed after the introduction of the Army Alpha were designed to reduce possible sex differences. As it became more apparent that males excel on some types of measures and females on others, and that all types of measures provide some information about general intelligence, psychologists designed general mental ability tests to include a balance of all types of measures. General intelligence measures such as the Stanford-Binet, and the Wechsler Adult Intelligence Scale (WAIS), were designed to avoid giving advantage to either sex (Tyler, 1965). Hence, male and female mean test scores on measures of general intelligence may differ slightly but these differences represent no practical significance.

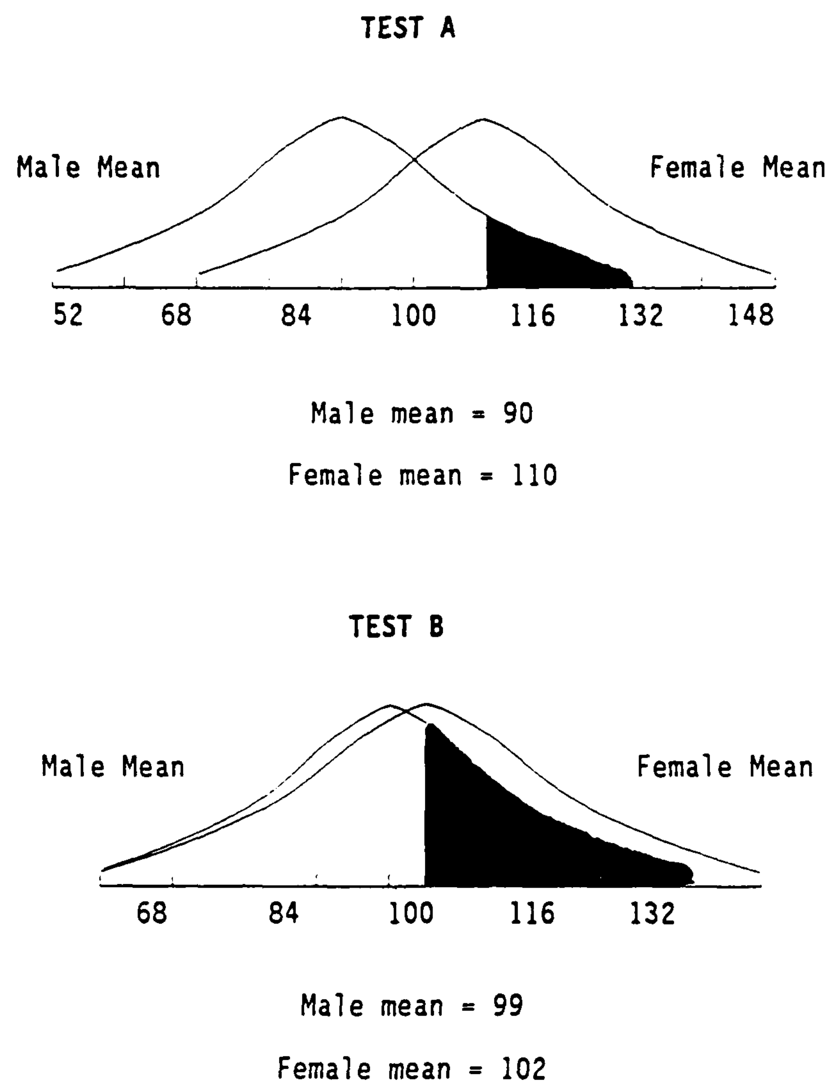


Figure 2. Overlap Between Male and Female Distributions for Two Tests with Statistically Significant Mean Score Differences

That is, given the amount of test score variance within each group, information about a person's gender provides little information about his or her level of measured general intelligence.

In utilizing a general cognitive ability measure such as the Armed Forces Qualification Test (AFQT) score[17] to select recruits for the Army, one would expect very small differences between males' and females' mean test score performance. In fact, one study, Profile of American Youth, indicated that nationally representative samples of males and females aged 18 to 23 years, differ very little in mean AFQT scores; the mean for males is 50.8, that for females is 49.5. (The standard deviation for the total sample is 28.03; total N = 25,409 [Department of Defense, 1982].)

Mean Score Differences Between Males and Females: Specific Cognitive Abilities

In the area of specific cognitive abilities, significant male-female test score differences do appear. To highlight these differences, we examine male-female differences on two multi-aptitude batteries--one designed for educational selection, the other for military selection and classification purposes. In this discussion, we mainly examine mean score differences for cognitive ability measures, although means for technical knowledge tests are presented.

Differential Aptitude Tests. Results for a study in which the Differential Aptitude test (DAT) was administered to over 5,000 male and 5,350 female 12th-grade students provide information about how the two groups differ on measures of cognitive ability constructs (Bennett et al., 1973). These data are presented in Table 17. Due to large sample sizes, all mean differences computed between males and females are statistically significant ($p < .01$). Close inspection provides information about the size of the differences and about the amount of overlap between the two distributions. For example, males score higher than females on the Verbal Reasoning, Numerical Ability, and Abstract Reasoning subtests, but these differences are slight; they represent an effect size of 0.13 or less. In terms of overlap, approximately 44 percent of the females score at or above the male mean. Thus, when scores on these tests are used to make selection decisions, very similar selection rates for males and females result.

[17]As reported earlier, the current AFQT score is derived from a composite of a person's score on four Armed Service Vocational Aptitude Battery (ASVAB) subtests--Word Knowledge, Arithmetic Reasoning, Paragraph Comprehension, and Mathematical Knowledge. At the time the study under discussion was conducted, the AFQT was computed using scores from Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Numerical Operations subtests.

Table 17

Twelfth-Grade Male and Female Mean Scores and Standard Deviations and Mean Effect Size Values for the Differential Aptitude Subtests

	Male (N = 5000+)		Female (N = 5350+)		Effect Size in SD Units
Subtest	Mean	SD	Mean	SD	
Verbal Reasoning[a]	31.1	12.2	30.5	12.3	.05
Numerical Reasoning[b]	24.9	9.8	23.7	9.2	.13
Abstract Reasoning[b]	35.8	10.1	34.9	10.0	.09
Clerical Speed and Accuracy[b]	45.8	11.8	51.6	11.9	-.50
Mechanical Reasoning[b]	50.6	10.6	41.1	10.0	.92
Space Relations[b]	34.3	13.0	30.9	11.9	.27
Spelling[b]	71.8	17.3	80.2	14.5	-.53
Language Usage[b]	33.8	11.4	38.3	10.9	-.40

Note: From Manual for the Differential Aptitude Test by G. K. Bennett, H. G. Seashore, and A. G. Wesman, 1973, New York: The Psychological Corporation. (Copyright 1973 by the Psychological Corporation.) Reprinted by permission.

[a]$p < .01$.
[b]$p < .001$.

Mean differences on the remaining measures would, however, yield more disproportionate selection rates for males and females. On Space Relations, which measures the ability to visualize a three-dimensional object from a two-dimensional display, males, on the average, obtain scores 0.27 standard deviation higher than females. Thus, only 40 percent of the females score at or above the male mean. Scores on Mechanical Reasoning indicate that males, in general, score nearly one standard deviation higher than females; only about 16 percent of the females score at or above the male mean on this measure.

Females, on the other hand, obtain higher mean scores than males on Clerical Speed and Accuracy, a measure of perceptual speed and accuracy. This represents a difference of one-half standard deviation or an effect size of 0.50. Thus, about 30 percent of the males score at or above the female mean on Clerical Speed and Accuracy. Similar effect size differences appear for the Spelling and Language Usage subtests (0.53 and 0.40 respectively).

According to these results on the DAT, males and females obtain fairly similar scores on measures of reasoning and numerical ability. Somewhat greater differences appear on measures of spatial ability and much greater differences appear on measures of mechanical aptitude, with males scoring higher than females on both. Although the differences are not as great as those for the mechanical aptitude, females obtain higher scores than males on measures of perceptual speed and accuracy.

<u>Armed Services Vocational Aptitude Battery</u>. Male-female mean test score differences have also been examined using the Armed Services Vocational Aptitude Battery (ASVAB). These data were obtained from a study entitled <u>Profile of American Youth</u> (Department of Defense, 1982) that was designed to examine the "cross-sectional character" of eligible military enlistees (Doering, Eitelberg, & Sellman, 1982). The sample includes over 9,100 young adults from ages 18 to 23 years and contains approximately equal numbers of males and females selected to be geographically representative of all youth throughout the United States. Male and female mean test scores for seven cognitive ability measures are provided in Table 18, along with mean scores for three technical knowledge tests.

Once again, because of sample size, all differences between male-female ASVAB subtest mean scores are statistically significant, except on the Word Knowledge subtest. Mean score differences on the Numerical Operations, Mathematics Knowledge, and Paragraph Comprehension subtests represent an effect size difference of 0.19 or less. The effect size difference on the Arithmetic Reasoning subtest is slightly larger ($\underline{d}$ = 0.28).

Greater mean score differences appear on the remaining measures. On the Mechanical Comprehension (and on Electronics Information) subtest, the male mean scores exceed the female mean scores by about 0.83 standard deviation. Females score higher than males on the Coding Speed subtest, reflecting an effect size difference of about 0.42. On the General Science and Auto/Shop Information subtests, male mean scores exceed those for females by 0.36 and 1.25 standard deviation units, respectively.

Table 18

ASVAB Subtest Scores of the 1980 Youth Population for Total Group, and Males and Females, and Mean Effect Size Values

ASVAB Subtest	TOTAL (N = 9,173) Mean	SD	MALES (N = 4,550) Mean	SD	FEMALES (N = 4,623) Mean	SD	Mean Effect Size In SD Units
Cognitive Abilities							
Arithmetic Reasoning[a]	50.3	10.25	51.7	10.47	48.9	9.82	.28
Word Knowledge[b]	50.8	10.05	50.8	10.32	50.9	9.77	-.01
Paragraph Comprehension[a]	51.5	9.66	50.6	10.03	52.4	9.18	-.19
Numerical Operations[a]	48.6	10.65	47.6	10.75	49.6	10.44	-.19
Coding Speed[a]	51.9	10.10	49.9	9.78	54.1	9.99	-.42
Mathematics Knowledge[a]	51.8	10.77	52.6	11.12	51.1	10.34	.14
Mechanical Comprehension[a]	47.6	9.55	51.2	9.73	43.9	7.79	.83
Technical Knowledge							
General Science[a]	49.6	9.69	51.3	10.09	47.9	8.94	.36
Auto and Shop Information[a]	46.3	9.92	51.4	9.77	40.9	6.75	1.25
Electronics Information[a]	47.6	9.55	51.5	9.73	43.9	7.79	.86

Note: From Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery. Washington, DC: Department of Defense (1982).

[a]Mean differences are significant at $p < .001$.
[b]Differences are statistically non significant.

Conclusions drawn from the ASVAB concerning male-female differences on cognitive ability measures are very similar to those from the DAT results. That is, mean score differences between males and females generally appear on most cognitive ability measures, although some of the differences are small. Males tend to perform slightly better than females on measures involving reasoning and mathematics ability, females perform slightly better than males on measures of reading comprehension, and both groups perform equally well on the word knowledge measure. The greatest differences between the two groups are on the speeded subtests, on which females score higher, and measures tapping mechanical comprehension, on which males score higher.

Memory. Male and female mean score differences also appear on measures of other cognitive ability constructs not included in the above two multi-aptitude test batteries. According to Tyler (1965), results from "most studies agree that females excel in rote memory" (p. 246). Measures of this construct require exact repetition of a group of digits or words immediately following the presentation of word-number pairs. Within this same construct area, females also score higher, on the average, than males on measures of visual memory, the ability to recall details, relationships between objects, or compass directions of objects located on a previously presented map.

According to Wilson and Vandenberg (1978), females score higher than males on measures requiring immediate visual memory (female mean of 15.7 vs. male mean of 15.2; total group SD = 3.0). For measures of delayed visual memory, females again obtain higher scores, on the average, than males (female mean of 12.3 vs. male mean of 11.6; total group SD = equals 3.7). Although the mean differences for both types of measures are statistically significant, they reflect an effect size of less than 0.20 for both measures (female N = 1,069, and male N = 1,027).

Perception. Male and female mean score differences also appear in the area of perception. To identify mean effect size differences on this ability construct, we refer to a detailed review designed to examine sex differences for both perceptual and spatial abilities (Sevy, 1982). This review represents a compilation of more than 50 years of research assessing mean differences between males and females. The reviewer included samples representing a wide variety of age ranges such as preschool, elementary, high school, and college students. From each study, Sevy identified the type of measure used, the cognitive ability construct assessed, the age or grade level of the sample, and the effect size difference. To quantitatively represent observed sex differences across all studies, Sevy used a meta-analytic method. Below we describe the tasks involved in measures of perception and then summarize the mean effect size observed for high school samples.

Measures of perception include tests designed to assess the ability to use some visual cues while ignoring others to identify figures or objects or to adjust objects to an upright or vertical position. This ability is also referred to as field independence. The first type of task is typically measured in a paper-and-pencil test that presents subjects with one or more simple figures or forms. After examining the forms, subjects are asked to identify one of the forms embedded in a complex figure. The Group Embedded Figures Test (Oltman, Raskin, & Witkin, 1971) and ETS Hidden Patterns

(Ekstrom et al., 1976) are two examples of this type of measure. Mean effect size differences for samples of high school students indicate that males generally score 0.31 standard deviation higher than females on the measure.

On field independence measures requiring apparatus, such as the Rod and Frame Test, males also outscore females (Witkin et al., 1954). In this measure the subject is asked to view a rod presented in an illuminated frame; no other features of the immediate environment are visible. With both frame and rod adjusted out of true vertical position, the subject is asked to adjust the rod to its true upright position. The subject able to locate true verticality, independent of the offsetting cues from the frame, is termed field independent. The subject who cannot separate the offsetting contextual cues of the frame in adjusting the rod is termed field dependent. The mean effect size computed across studies for high school students is 0.48. Across both types of measures of field independence, then, 30 to 40 percent of the females obtain scores at or above the male mean.

Spatial Abilities. One final cognitive ability construct requiring closer examination involves spatial ability. Because the construct includes several types of spatial tasks, it is important to examine male-female mean score differences on each. As noted above, Sevy (1982) also examined mean effect size differences for several types of spatial ability tests. Results from this meta-analysis are summarized in the following discussion for each separate spatial ability task.

Space visualization involves the ability to mentally manipulate the components of two- or three-dimensional figures into different arrangements. Numerous paper-and-pencil measures have been developed to assess this ability. For example, in one measure the task involves visualizing the appearance of an object assembled from a number of separate parts (Flanagan Industrial Tests - Assembly, Flanagan, 1975). In another, subjects are asked to visualize objects in three-dimensional space in order to count the number of objects adjacent to a target object (Employee Aptitude Survey - Space Visualization, Ruch & Ruch, 1980). On these types of measures, the mean effect size computed across studies which included high school students indicates that males generally score 0.34 standard deviation higher than females. In terms of overlap, 37 percent of the females score at or above the male mean.

Measures requiring two-dimensional spatial rotation present subjects with standard figures such as cards or flags. Test items include figures that are the same as the standard figures except that they are rotated, or figures that are different from the standard figures because they are inverted. Subjects are asked to compare test figures with standard figures to determine whether they are the same or different. Examples of these types of measures include Thurstone's Flags (Thurstone & Jeffrey, 1979) and ETS Card Rotations (Ekstrom et al., 1976). On these types of measures, male high school students outscore female high school students by 0.42 standard deviation, indicating that 34 percent of the females score at or above the male mean.

Three-dimensional spatial rotation measures include tasks very similar to those required in two-dimensional spatial rotation measures, but the task here requires subjects to visualize a three-dimensional object depicted in two-dimensional space. Subjects must mentally rotate the target object to determine whether test objects are the same as the standard object. This type of task is required in the Shepard-Metzler Mental Rotation Test (Wilson & Vandenberg, 1978). Across studies that include high school students, the mean effect size of 0.92 indicates that only 19 percent of the females obtain scores at or above the male mean. Across these three spatial ability subcomponents, males score, on the average, higher than females, but mean effect size differs with the type of task involved. On measures requiring only visualization of two- or three-dimensional objects, male-female differences are smaller than differences observed on measures requiring both visualization and rotation of two-dimensional objects. Even greater sex differences appear for measures requiring visualization and rotation of three-dimensional objects.

Concerning the spatial orientation construct, male-female differences are less clear. In the U.S. Army Air Force Aviation Psychology Program, mean scores for several spatial orientation measures were provided separately for male and female pilot trainees (Guilford & Lacey, 1947). Although the samples represent highly select groups, data from this study provide information about how males and females differ. For example, one measure, Instrument Comprehension, involves two slightly different types of tasks. In Part One, subjects are asked to review airplane readings on six instruments or dials and then select the correct written description of the plane's position. Part Two requires subjects to examine two airplane instruments and then select the correct pictorial representation of the plane's position. Mean test scores computed separately for each part revealed that females and males obtain approximately equal scores on Part One (male mean = 9.71 and female mean = 9.17; standard deviation for the total group is 3.20), whereas males outscore females on Part Two (male mean = 32.75, and female mean = 25.05; standard deviation for the total group equals 10.29). This represents an effect size of 0.17 for Part One and 0.75 for Part Two. In terms of overlap, 43 percent of the females score at or above the male mean on Part One, whereas only 23 percent of the females score at or above the male mean on Part Two. From these data, it is clear that very similar types of tests designed to tap spatial orientation may actually measure different constructs. Items contained in Part One of Instrument Comprehension appear to include a combination of spatial orientation, reading comprehension, and verbal ability, whereas items in Part Two appear to measure only spatial orientation.

Results from Sevy's (1982) review of the literature also present problems in drawing conclusions about sex differences on spatial orientation. Only nine studies designed to assess mean sex differences on measures of spatial orientation could be located. When mean effect sizes are examined by age group, it appears that differences are smaller for high school samples (0.39) than for college samples (0.85). Although these effect size differences may be due to sample differences, they are more likely due to differences between measures administered to the two groups.

Summary

Overall, then, measures of cognitive ability constructs typically yield mean score differences between males and females, but for many constructs these differences are very small. Mean score differences expected between males and females on the cognitive ability constructs discussed above are summarized in Table 19. As shown, measures of general intelligence and verbal ability yield only inconsequential mean score differences between males and females. On measures of reading comprehension and memory, females as a group score slightly higher than males, while on measures of numerical ability, reasoning, and field independence, males score slightly higher than females. Greater mean score differences appear on measures of perceptual speed and accuracy, with females scoring on the average about 0.40 to 0.50 standard deviation unit higher than males. On measures of spatial orientation (excluding items that tap reading comprehension and verbal ability), males outscore females by about 0.39 to 0.85 standard deviation unit. Measures assessing spatial visualization and mental rotation abilities yield varying mean effect size differences, ranging from 0.34 to 0.92 standard deviation unit. On measures of mechanical ability and related technical knowledge tests, male and female mean scores differ by about one standard deviation.

Consistent and reliable mean effect size differences observed between males and females on cognitive ability measures influence selection decisions. Earlier, we provided an example of two measures yielding different effect sizes and resulting in different selection rates for males and females. Thus, reliable effect size differences suggest that disproportionate selection of one group over another may occur. The actual selection ratio for males and females will vary with mean effect size, differences between male and female test score variances, and the test cut-off score. For example, if males as a group obtain scores 0.30 standard deviation higher than females as a group, and the two groups have equal test score variances, then the selection ratio of males to females with a cut-off score set at the total group mean is 1.25 (Sevy, 1988).

According to Federal guidelines, this selection ratio for two groups constitutes adverse impact. (Adverse impact is defined and discussed in detail later in this subsection.) Although the male-female selection ratio may be equalized by lowering the cut-off score, with extremely large effect size differences (e.g., 0.80 or greater), very low cut-off scores continue to yield disproportionate selection rates. For example, for a test yielding an effect size of 0.80, the cut-off score must be set at -2.0 or -2.5 standard deviations below the total group mean to ensure equal selection rates for males and females. Thus, on measures yielding consistent and large effect size differences, such as measures of three-dimensional spatial rotation or mechanical aptitude, disproportionate selection rates occur between males and females even with very low cut-off scores.

GROUP DIFFERENCES IN COGNITIVE ABILITY TEST PERFORMANCE: RACE AND ETHNIC GROUP DIFFERENCES

The notion that persons of different races or belonging to different ethnic groups vary on general intelligence measures has been postulated and under study for well over a century. For example, Sir Francis Galton, a pioneer in the field of differential psychology, suggested in 1869 that

Table 19

Summary of Male - Female Mean Score Differences in Cognitive Ability Constructs

Construct	Higher Scoring Group	Amount of Difference (in SD units)
General Intelligence	Equal	--
Verbal Ability	Equal	--
Reading Comprehension	Females	.19
Memory	Females	.20
Numerical Ability	Males	.13
Reasoning	Males	.13 to .27
Perception (Field Independence)	Males	.27 to .34
Perceptual Speed and Accuracy	Females	.40 to .50
Spatial Orientation (Excluding Reading Comprehension and Verbal Ability)	Males	.39 to .85
Spatial Ability (Visualization and Mental Rotation)	Males	.34 to .92
Mechanical Ability	Males	.92 to 1.00

different races could be ordered along a continuum of high versus low intelligence. Subsequent research in this area has provided insight about mean group differences on measures of general intelligence, and group strengths and weaknesses on specific cognitive ability measures. In this part, we review the data related to the comparison of mean intelligence scores and cognitive ability test scores across race and ethnic subgroups, and then examine the cognitive ability profiles within each group.

Methodological Issues

Before examining race and ethnic subgroup differences, it is important to identify methodological pitfalls that may influence these comparisons. Previously we identified three factors that Anastasi (1937) argues have an impact on observed differences between male and female mean test scores. These same factors may influence or confound results from race and ethnic subgroup comparisons.

The first factor, selective attrition, implies that different rates of elimination from the subject pool for different race and ethnic subgroups result in test score distributions and mean test scores that do not reflect the true values for the target subgroup populations (e.g., high school age youth).

The second factor, socialization, may also influence observed mean differences between race and ethnic subgroups. This would occur if a subgroup's membership is related to opportunity to practice cognitive tasks both in the home and at school. For example, it is widely accepted that conditions for experiencing intellectual stimulation in the home and in school are far more prevalent in upper than in lower class environments. Thus, when a larger percentage of one race or ethnic group than another race or ethnic group is found in lower socioeconomic status (SES) environments, it is not surprising that members from the lower SES subgroup obtain relatively low mean scores on cognitive ability measures.[18]

The third factor, the effects of sample size, makes it important to include fairly large, representative samples from each target subgroup to reduce sampling error and to ensure reliability of observed mean differences. A related issue involved in race and ethnic group comparisons is sample selection. Findings from several studies indicate that in addition to between-subgroup differences, within-group differences appear by region, socioeconomic level, and locale (Anastasi, 1937; Jensen, 1980; Willerman, 1979; Yerkes, 1921). Thus, sampling from a single region, SES group, or locale yields samples that are not exactly representative of one or more race or ethnic groups. Sample selection, then, must take into account variables that may confound or cloud true subgroup differences.

[18]The notion that socialization differences between racial and ethnic subgroups confound mean score comparisons on cognitive ability tests may be analogous to cultural bias issues in testing. The validity of cultural bias arguments is examined in the following section.

A final consideration in comparing race or ethnic subgroups involves race identification. According to Anastasi, race is a biological or genetic term; thus physical features are often used to identify racial heritage. Problems arise with this usage, however. For example, in our culture a racial identification is often based on skin color alone rather than on racial heritage. For some, this results in a genetically erroneous classification when parentage is not considered (e.g., if three out of four grandparents are white yet a person is classified as black). In our culture, then, race identification is determined more by social acceptance than by true racial heritage (Willerman, 1979).

Other physical features used to identify racial heritage include pigmentation of eyes, hair color, hair texture, gross body dimensions such as stature, or facial and cranial measurements. Relying on these features to determine racial heritage also invokes problems because of the wide variability within any one group and because of the amount of overlap between groups. It is difficult, then, if not impossible, to classify persons into "pure" racial groups. For the most part, researchers rely on participants to indicate the race or ethnic group with which they identify.

Racial or ethnic group identification does not represent a well-defined or distinct classification system. Thus, when cognitive ability mean test scores are compared by race and ethnic subgroup, several unavoidable factors, such as selective attrition, socialization, and problems with race identification, may cloud or confound results. These and other problems may be circumvented by relying on subgroup samples that are sufficiently large and representative of target subgroup populations. In the following discussion we present results of race and ethnic subgroup comparisons from studies using fairly large, representative samples.

Race and Ethnic Subgroup Mean Score Differences: General Intelligence

During World War I, approximately 1.7 million men were assessed using the Army Alpha, Army Beta, or individually administered intelligence tests. Results from this large-scale administration permitted the examination of mean score differences by nationality on measures designed to tap general intelligence. Yerkes (1921) compared group median scores for more than 12,500 white foreign-born draftees representing 16 European countries. Results from this analysis indicated that, compared to the median value for native-born, white draftees, the English and Scottish obtained higher median scores; Germans, Irish, and Scandinavians obtained median scores approximately equal to those of native-born whites; and Russians and Southern Europeans obtained the lowest median scores. Although not included in this particular analysis, native-born black draftees typically obtained scores similar to the Southern Europeans.

These data must be viewed with caution, because as Anastasi (1937) notes, immigrant groups are not representative of the home population. Reasons for immigrating may vary from one country to the next. Thus, immigrants from one country may represent a random sample of the home population while immigrants from another may represent a more select group. Further, length of time spent in the United States influences mean test scores. For example, analyses of general intelligence test scores obtained from a sample of white foreign-born draftees indicated that those who had been in the country longer, 20 years or more, obtained higher scores than those who had been in country 5 years or less (i.e., 13.70 versus 11.30, respectively [Yerkes, 1921]).

Jensen (1980) summarized a more recent report providing data on a representative sample of subjects (Coleman et al., 1966a). In this study, a nationwide sample of more than 645,000 students in grades 1, 3, 6, 9, and 12 were tested on verbal and nonverbal aptitude tests and scholastic achievement tests. The aptitude tests are from standard group tests of verbal and nonverbal intelligence and contain items such as picture vocabulary, picture association, classification, sentence completion, and figural and verbal analogies. The achievement tests measure reading comprehension and mathematics achievement. Results from this study are reported in Table 20 for black, Mexican, American Indian, and Oriental students in grade 12. Note that these data are reported as mean effect size differences between white and minority group means.

Table 20

Difference Between White Majority and Minority Group Means Expressed in Standard Deviation Units[a] (12th Grade Level Only)

Test	Minority Group			
	Black	Mexican	American Indian	Oriental
Verbal I.Q.	1.24	0.91	0.93	0.28
Nonverbal I.Q.	1.31	0.82	0.57	-0.04
Reading Comprehension	1.05	0.85	0.84	0.35
Mathematic Achievement	1.13	0.72	0.70	0.07

Note: Calculated from Coleman et al. (1966b), presented in Jensen, "Bias in Mental Health Testing," 1980, p. 479. New York: The Free Press. (Copyright 1980 by the Free Press.) Reprinted by permission.

[a]Raw score means and standard deviations were used (mean effect size = white mean - minority mean/white standard deviation).

According to these data, the Oriental group mean on the verbal I.Q. differs only slightly from the white mean and virtually no differences exist on the nonverbal I.Q. Both Mexican and American Indian group means on the verbal I.Q. differ from the white mean by slightly less than one standard deviation. On the nonverbal I.Q., the American Indian group mean differs from the white mean by about 0.50 standard deviation unit and the Mexican mean differs by 0.82 standard deviation unit. On both verbal and nonverbal I.Q. measures, the mean for blacks differs from the mean for whites by over one standard deviation.

Subgroup mean differences on the reading comprehension test are very similar to those on the verbal I.Q. The Oriental group mean is about 0.35 standard deviation unit lower than the white; the Mexican and the Indian group means are about .84 standard deviation unit lower than the white; and the black is one standard deviation lower.

On the mathematic achievement test, the Oriental group mean is virtually the same as the white group mean, the Mexican and American Indian group means are about 0.72 standard deviation unit lower than the white mean, and the black group mean is over one standard deviation below the white mean.

Although the amount of the mean score difference varies across the four types of measures, the same pattern emerges in each. Whites and Orientals obtain approximately equal group means; Mexicans and American Indians obtain mean scores about 0.50 to 0.80 standard deviation unit below whites, and blacks obtain means one standard deviation below the white mean.

Race and Ethnic Subgroup Mean Score Differences: Specific Cognitive Abilities

Another study providing more details about race and ethnic group differences on several cognitive ability measures and technical knowledge measures is the Profile of American Youth (Department of Defense, 1982). In this study, ASVAB subtest scores were obtained for a nationally representative sample of white, black, and Hispanic youth. Mean subtest scores for cognitive ability and technical knowledge ASVAB subtests are provided in Table 21 along with mean scores for the total group.

Results from this study indicate that the mean score for whites is consistently and significantly higher across all ASVAB subtests than the mean scores for blacks and Hispanics (see Table 22). For cognitive abilities, the greatest difference between black and white mean test scores appears on measures of Word Knowledge and Mechanical Comprehension (mean effect size equals 1.20 or greater). The subtest yielding the smallest difference between black and white mean scores is the Mathematics Knowledge subtest (i.e., 0.88 standard deviation unit).

Hispanics' mean subtest scores differ the most from whites' on the Word Knowledge test (1.00 standard deviation unit). This group differs from whites the least on Numerical Operations, Coding Speed, and Mathematics Knowledge (mean effect size equals 0.73 or less).

Table 21

Armed Services Vocational Aptitude Battery (ASVAB) Mean Subtest Scores for Total Group and by Racial/Ethnic Group[a]

	Racial/Ethnic Group			
ASVAB Subtest	Total (N = 9,173)	White (N = 5,533)	Black (N = 2,298)	Hispanic (N = 1,342)
Cognitive Abilities				
Arithmetic Reasoning	50.3 (10.25)	52.3 (9.77)	41.6 (7.48)	44.0 (9.18)
Word Knowledge	50.8 (10.05)	53.0 (8.47)	41.7 (10.84)	43.9 (11.18)
Paragraph Comprehension	51.5 (9.66)	53.3 (8.41)	43.5 (10.52)	45.2 (11.26)
Numerical Operations	48.6 (10.65)	50.3 (9.74)	40.7 (11.05)	43.2 (11.42)
Coding Speed	51.9 (10.10)	53.5 (9.40)	44.4 (9.91)	47.7 (10.60)
Mathematics Knowledge	51.8 (10.77)	53.5 (10.54)	44.7 (8.36)	45.9 (9.93)
Mechanical Comprehension	47.6 (9.55)	49.4 (9.05)	39.3 (6.80)	41.8 (9.10)
Technical Knowledge				
General Science	49.6 (9.69)	51.7 (8.60)	40.9 (8.94)	42.6 (10.67)
Auto and Shop Information	46.3 (9.92)	48.2 (9.29)	37.4 (7.34)	40.5 (9.99)
Electronics Information	48.0 (9.86)	50.0 (9.05)	39.2 (8.19)	41.4 (10.05)

Note: From Profile of American Youth, Department of Defense (1982).

[a]Standard deviations are shown in parentheses.

Table 22

Differences Between Race/Ethnic Group Means on ASVAB Cognitive Tests, Expressed in Standard Deviation Units[a]

ASVAB Cognitive Ability Subtest	Subgroup Pairs		
	Black-White	Hispanic-White	Black-Hispanic
Arithmetic Reasoning	1.17	.86	.29
Word Knowledge	1.23	1.00	.20
Paragraph Comprehension	1.08	.90	.16
Numerical Operations	.95	.71	.22
Coding Speed	.95	.60	.32
Mathematics Knowledge	.88	.73	.13
Mechanical Comprehension	1.20	.84	.32

Note: Computed from data provided in Profile of American Youth, Department of Defense (1982).

[a]Mean effect size in standard deviation units equals higher mean - lower mean/pooled standard deviation.

Blacks' and Hispanics' mean subtest scores do not differ greatly from each other in terms of mean effect size. The greatest differences appear on the Mechanical Comprehension and Coding Speed subtests (0.32 standard deviation units). The two groups differ the least on Mathematics Knowledge (0.13 standard deviation unit).

Results from this study lend support to the results reported by Coleman and colleagues. That is, in a nationally representative population of high school youth, whites on the average score higher than Hispanics, who, in turn, score slightly higher than blacks on measures of cognitive abilities. These same conclusions hold when considering mean scores on a general measure of intelligence, the AFQT, computed from scores on four ASVAB subtests (white group mean, 55.9; Hispanic group mean, 31.5; black group mean, 24.3; and total mean, 50.1).

Race and Ethnic Subgroups: Within-Group Profiles

Another way to view race and ethnic group differences is to examine the cognitive ability profiles within each group. Most of the literature providing such data has dealt with preschool or elementary school children. Thus, the following summary describing within-group differences is based upon research that includes young children. Much of the data is provided on a broad grouping of cognitive abilities. Therefore, the profiles offer only very crude descriptions of race and ethnic within-group differences across measures of verbal, reasoning, numerical, and spatial abilities.

> The Hispanic profile appears relatively flat, with the lowest scores on verbal ability measures, and slightly higher scores on reasoning ability measures. The highest scores (albeit only slightly higher) appear on numerical and spatial measures (Willerman, 1979).
>
> The Oriental profile (includes Chinese and Japanese) indicates this group scores lowest on measures of verbal ability and much higher on measures of numerical, reasoning, and spatial ability (Willerman, 1979).
>
> American Indians appear to score lowest on verbal and reasoning ability measures and highest on measures of spatial ability (Tyler, 1965).
>
> Blacks tend to perform best on verbal ability measures, slightly lower on spatial and reasoning measures, and lowest on numerical ability measures (Willerman, 1979).

Summary

The purpose of this review has been to identify the degree to which race and ethnic subgroups differ on general intelligence and cognitive ability measures. Studies of general intelligence show that, on the average, American Indians, Hispanics, and blacks score lower than whites. Similar group differences appear on measures of specific cognitive abilities. An important point to make here concerns the amount of overlap between race and ethnic subgroups. Even with observed mean score differences as great as one standard deviation, one cannot predict a test score for a single individual based on knowledge of race or ethnic group.

Because of these observed group differences, the question of possible bias in testing has been raised. In the following subsection, we review these concerns and the evidence accumulated to address them.

ISSUES OF TEST BIAS AND TEST FAIRNESS

Because mean scores for majority and minority group members differ on measures of general intelligence and on measures of specific cognitive abilities, considerable attention has been given to investigating possibilities that these differences may stem from possible bias in test measures.

The concept of test bias can be viewed in several ways. First, from a test construction view, content may provide an advantage to one subgroup over another. Second, from a statistical view, test score meaning or interpretation may vary for minority or majority subgroup members. In other words, interpreting test scores via predictive validity coefficients may provide useful information about potential for success for one subgroup but not for another subgroup. Differential validity (differences between subgroup validity coefficients) and differential prediction (differences between subgroup regression slopes, intercepts, and standard errors of estimate) indicate whether or not test scores *have the same meaning for different* subgroups.

In this part we examine the evidence related to test content bias, differential validity, and differential prediction. In addition, we review and evaluate various approaches (or "test fairness models") which have been proposed as possible solutions to problems of bias in the use of tests for selection.

Definition of Terms

Before examining these issues, we provide some definitions of test bias and test fairness terms as used in this review. In large part, these definitions are derived from Jensen's (1980) review of test bias.

As noted above, the existence of test bias may be determined by both subjective and objective (i.e., statistical) procedures. These terms are defined as follows:

> Content Bias indicates bias occurs if items contained in the test give one subgroup an advantage over another subgroup because of greater opportunity to learn or acquire information. Content bias is generally determined by subjective means.

Differential Validity indicates a test is biased if the validity coefficient for one subgroup differs significantly from the coefficient of another subgroup.

Differential Prediction indicates a test is biased if for two subgroups statistically significant differences appear between regression equation slopes, or regression equation intercepts, or standard errors of estimate of the regression lines.

Test Fairness refers to the way in which scores on a test, whether biased or unbiased, are used in practical applications. A biased test may be used fairly and an unbiased test may be used unfairly. Determination of test fairness involves social, political, legal, and moral issues. Although many complex statistical models have been developed to ensure test fairness, the exact model or approach used rests upon legal, philosophic, and practical considerations.

Finally, to clarify the meaning of subgroup terms majority and minority, the majority group is defined as (a) the larger of two groups in the total population, and (b) the group on which the test was primarily standardized.

Test Content Bias

Cultural Bias Theory. According to Jensen (1980), Binet was the first to express concern about test content bias. Binet recognized that intelligence measurement presupposes common language and common cultural and background experiences. If test items are not carefully sampled, measurement of intelligence can be biased or contaminated for persons or groups with atypical educational or cultural experiences. As they were developing tests early in this century, Binet and Simon attempted to avoid this bias by excluding items tapping specific knowledge acquired at home or at school. In this country, test content bias issues can be traced back at least six decades to the 1920s, when intelligence was considered immutable; hence, intelligence test scores, when low, limit individuals' opportunities (Carroll, 1982).

In seeking to identify possible origins of bias in test content, it has been suggested that, because most tests are developed by white, middle-class persons, test items reflect information acquired in a white, middle-class environment. It is pointed out that such measures, when used to assess members of other subgroups, may yield erroneous scores for those subgroups because their environment may not have afforded equivalent learning opportunities. Thus, measures of general intelligence and of specific cognitive abilities could be biased against subgroups other than the white, middle-class.

Several reasons have been advanced for the occurrence of bias and for problems with continued use of intelligence tests. For example, Haggard (1954) stated that children of low socioeconomic status are handicapped in taking written tests because of reading difficulties; thus, orally administered tests should be given to these children. Katz and Greenbaum (1963) argued that bias occurs because of differential motivation to perform well on tests. Fine (1975) and Daniels (1976) stated that biased measures of intelligence perpetuate the inferior status of ethnic-minority and low socioeconomic groups. Williams (1970) contended that it is unclear exactly what intelligence tests are measuring; therefore, he called for a moratorium on testing until more is known about their suitability for black students.

Publications on this issue often do not provide details about what constitutes content bias or cultural bias in test items. Perhaps the clearest definition of this concept is provided by Eells, Davis, Havighurst, Herrick, and Tyler (1951):

> By <u>cultural bias in test items</u> is meant differences in the extent to which the child being tested has had the opportunity to know and become familiar with the specific subject matter or specific process required by the test item. If a test item requires, for example, familiarity with symphony instruments, those children who have opportunity to attend symphony concerts frequently will presumably be able to answer the question more readily than those children who have never seen a symphony orchestra. To the extent that intelligence-test items are drawn from cultural materials of this sort, with which high [socioeconomic] status pupils have more opportunity for familiarity, status differences in I.Q.'s will be expected. (p. 58)

Eells and associates also indicated how cultural bias operates with other variables to create differences between majority and minority mean test scores:

> Both genetic and developmental factors are presumed to determine the actual intelligence of the child as it might be evidenced in thinking clearly and in solving appropriate problems in real-life situations. . . . [C]ultural bias in test items, test motivation, and test work habits or test skills, on the other hand, are oriented toward the test situation as such and are assumed to affect the pupil's ability to score well on the test but not to affect materially his ability to think clearly and to solve appropriate problems in real life situations. (p. 58)

Specific elements of the problem of test content bias have not yet been defined in sufficient detail for research purposes. For example, variables operating in different cultural environments that create subgroup differences on general intelligence tests have not been identified. Although one can readily isolate features that distinguish low from middle or high socioeconomic environments, the circumstances that actually produce subgroup differences have not been clearly or operationally defined. An example of a variable that characterizes differences between subgroup cultures, socialization, was discussed earlier in this section; it may include such things as parental, peer, and teacher encouragement to achieve in academic pursuits, and materials available at home or in school that stimulate cognitive development. Variables such as these, if measured in quantifiable form, may provide a more informative picture of cultural differences that produce subgroup differences on general intelligence and cognitive ability tests.

The question of cultural bias may be extended to other personal characteristics, such as temperament and vocational interests. Following the culture bias reasoning, different cultures or environments afford different standards of correct or deviant behavior and different opportunities for vocational experiences. Thus, fairly large mean score differences on temperament and vocational interest measures would be expected between members of different socioeconomic status subgroups and between race and ethnic subgroups (M. D. Dunnette, personal communication, 1984). Evidence for large race or ethnic subgroup differences on these types of measures is somewhat mixed, but not many differences appear. For example, on vocational interest measures, blacks and whites appear to differ very little. A review of black and white mean score differences in temperament reveals large differences on only one scale; mean score differences on the remaining scales reviewed are smaller and less consistent (Kamp & Hough, 1987). Thus, it is unclear why some personal characteristics, such as cognitive abilities, are influenced by cultural differences between race and ethnic subgroups, while other characteristics, such as temperaments and vocational interests, are not influenced or are influenced to a much lesser degree by these cultural differences.

A component needed to assess cultural bias involves criteria for judging test content. Typically, cultural bias in items is assessed by a panel of judges or experts who rely on their own definition of cultural bias. According to Jensen (1980), items that involve scholastic or "bookish" vocabulary or knowledge of fine arts, or items that reflect the values of the white middle class are judged to be culturally biased or culture-bound, and thus unfair to non-whites or persons of low socioeconomic status.

Examples of culturally biased items or items that tap information potentially unfamiliar to some subgroups can be found as far back as the Army alpha information test (Yerkes, 1921):

1. The knight engine is used in the:

   a) Packard b) Lozier c) Stearns d) Pierce Arrow

2. The Pierce Arrow car is made in:

   a) Buffalo b) Detroit c) Toledo d) Flint

3. An air-cooled engine is used in the:

   a) Buick b) Packard c) Franklin d) Ford

Knowledge of the correct responses might have been a function not only of interest in cars and engines, but, even more, of opportunity to be familiar with them in the early years of the automobile era. This type of item could have resulted in mean test score differences between urban and rural samples.

Removing items that give advantage to one group over another might be expected to reduce differences between subgroup mean test scores.

The focus of research on this issue, then, has been on the verbal components of most general intelligence measures because these types of items represent information acquired from daily living experiences in a typical white, middle-class environment. Opportunities to acquire this information are not equal for members of different subgroups (e.g., minorities, lower socioeconomic status persons). The converse of this theory would argue that all groups have equal opportunity to experience and acquire information necessary to learn answers to such questions.

Cultural Bias Theory: Empirical Evidence. Numerous studies have been conducted to explore the effects of culturally biased test items on mean score differences of blacks and whites. For example, Jensen (1980) described a series of studies conducted by McGurk (1953a, 1953b, 1967) intended to test the cultural bias theory. McGurk examined items from several general intelligence measures, such as the Otis test and the American Council on Education (ACE) test. A panel of 78 judges, including experts in the areas of psychology, sociology, and counseling, were asked to classify each of 226 general intelligence test items as (a) least culturally biased, (b) neutral, and (c) most culturally biased.

Items judged to be "most" culturally biased (103 items) and items evaluated as "least" culturally biased (81 items) by at least 50 percent of the judges were retained and then administered to a sample of 90 high school seniors, including both blacks and whites. From these data, difficulty levels on the most culturally biased were matched with least culturally biased items, yielding 37 pairs of items matched on difficulty.

Resulting items were administered to seniors in 14 high schools located in Pennsylvania and New Jersey (N = 2,630 whites and 233 blacks). For each student, three scores were computed: (a) score for items judged to be most culturally biased; (b) score for those items evaluated as least culturally biased, and, (c) total test score on most and least items combined. Means computed separately for blacks and whites on the three test scores--most, least, and total--were then compared.

Results indicated that for total test score, blacks as a group scored lower than whites by 0.50 standard deviation unit. For the two subtest scores, blacks' scores averaged 0.30 standard deviation unit lower than whites on the most culturally biased test score, and 0.58 standard deviation unit lower than whites on the least culturally biased test score. According to these data, tests containing items rated as most culturally biased yielded smaller mean differences between blacks and whites than tests containing items judged to be least culturally biased.

McGurk (1975) also reviewed literature reporting mean scores for blacks and whites on verbal and nonverbal measures of general intelligence. His rationale for comparing mean scores on verbal and nonverbal measures stemmed from culture bias theory (e.g., that cultural differences in opportunity to practice verbal tasks, or to become familiar with terms included in tests, produce observed differences in subgroup mean scores; hence, subgroups should differ less on nonverbal intelligence measures [Haggard, 1954]). In this study, McGurk computed mean effect size differences between mean scores of blacks and whites for both types of tests. Results showed greater overlap between groups on verbal measures than on nonverbal measures, suggesting that blacks and whites differed less, on the average, on tests judged to be more culturally biased than on tests thought to reduce cultural bias by eliminating verbal ability requirements.

Davis and Eells (1953) addressed the cultural bias issue by developing a test designed to reduce differences in subgroup mean scores. These researchers constructed a measure that would reduce motivational differences between lower and upper socioeconomic status, reduce reading requirements that pose greater difficulty for SES children, and assess information equally familiar to both groups. Information gleaned from a series of interviews with educators and sociologists familiar with characteristics of family living and child rearing at different socioeconomic status levels and from systematic observation of children in free-time activities was used to develop the Davis-Eells Test of General Intelligence or Problem Solving Ability. This test is composed entirely of cartoons or pictures, involves no reading, and is described to children as a game, to increase interest and motivation. For each item, an administrator asks children to examine a picture or series of pictures and (a) identify from among three pictures the best way to perform a task such as how to put in a new light bulb, (b) identify the solution to a pictorial analogy problem such as: "glove is to hand as sock is to ____ [foot]," and (c) select the best description of a picture (the administrator reads aloud three descriptive statements).

Jensen (1980) reported that in a majority of studies using the Davis-Eells test, mean test score differences between lower socioeconomic and middle-class children appeared to the same degree and in the same direction as differences in conventional intelligence tests (e.g., Angelino & Shedd, 1955; Coleman & Ward, 1955; Fowler, 1957; Noll, 1958). Blacks obtained slightly lower mean I.Q. scores on this measure than on other more commonly used general intelligence tests such as the California Test of Mental Maturity (Ludlow, 1956). Thus, the Davis-Eells games, although designed to reduce the advantage for white, middle-class children, yield mean score differences similar to conventional intelligence test score differences between social class groups and between blacks and whites.

Williams (1972) designed a test to demonstrate reverse cultural bias on a measure of general verbal intelligence or knowledge acquired through daily living experience. The Black Intelligence Test of Cultural Homogeneity (BITCH) was designed to assess specialized vocabulary peculiar to the black culture. Jensen (1980) provided example items typical of the content of the BITCH (p. 680):

1. The Bump

   (a) A result of a forceful blow
   (b) A suit
   (c) A car
   *(d) A dance

2. Running a game

   (a) Writing a bad check
   (b) Looking at something
   (c) Directing a contest
   *(d) Getting what one wants

---

* Correct answer.

Subgroup mean score differences on the measure are as intended; that is, mean scores for blacks and whites on this test are virtually non-overlapping, with blacks scoring much higher than whites. The measure, however, fails to correlate with scores on traditional intelligence measures (Arvey, 1979). Its predictive validity in educational or occupational settings is yet to be assessed (Jensen, 1980).

The BITCH represents a measure designed to assess information acquired from cultural experiences that differ from white, middle-class experiences. Mean scores for blacks and whites on this measure demonstrate that test items, when constructed to tap information more familiar to one subgroup than another, produce large and significant subgroup differences. This suggests that potential test content bias or cultural bias in general intelligence test items is a genuine concern. Test scores for blacks and whites on the BITCH, however, do not provide the same information as scores on traditional intelligence measures. Scores on the BITCH do not correlate with these measures. Traditional measures of intelligence, although potentially biased, do provide meaningful information about individuals' potential for success in our culture as a whole.

Attempts to explain subgroup mean score differences--or, more specifically, black and white mean differences--on traditional measures of intelligence by using test content bias arguments fail to produce results predicted by advocates (Arvey, 1972). For example, advocates contend that motivational differences in opportunity to acquire information, and reading comprehension and verbal ability differences between subgroups account for differences on intelligence tests. Research reviewed here, however, suggests that the above factors, do not, in fact, explain the mean score differences. Cultural bias arguments as currently postulated do not help us to understand why race, ethnic, and socioeconomic subgroup differences appear on measures of general intelligence or on measures of specific cognitive abilities.

Another way to view the cultural bias question is to follow it through to its logical conclusion: Race and ethnic subgroup differences will disappear if cultural or environmental experiences for the subgroups are equalized. Tyler (1965) described two studies in which blacks and whites appear to have had comparable environmental experiences. These studies are discussed below.

Tanser (1939) conducted a study that included black school children whose ancestors moved to Kent County, Ontario, prior to the Civil War period. In this particular community, white and black children had attended the same schools since 1890. Thus, at some level black and white school children had similar educational experiences. The entire population of black students in grades one through eight from one urban and seven rural schools participated in this study, along with white students from the same schools. Tanser administered both verbal and nonverbal intelligence tests to these students. Mean scores for blacks and whites appear in Table 23. According to these results, black and white mean scores differ by 15 to 19 points on both verbal and nonverbal tests.

Table 23

Black and White Mean Score Differences Reported in Tanser Studies

Measure	White			Black		
	N	Mean	SD	N	Mean	SD
National Intelligence Test	386	103.6	16.5	103	89.2	15.9
Pintner Non-Language Test	387	110.9	19.0	102	95.3	13.3
Pintner-Cunningham Primary Test	155	97.6	--	54	82.8	--
Pintner-Paterson Performance Test	211	109.6	22.4	162	91.0	19.0

Summarized from Tyler, 1965.

Note: From The Settlement of Negroes in Kent County, Ontario, by H. A. Tanser (1939), Chatham, Ontario: Shepherd Publishing Company.

Bruce (1940) conducted a study that included black and white school children from low-income regions of the South. This study complements Tanser's work by focusing on a sample in which all subjects are from the lower end of the economic continuum. Included in the sample were white and black school children ranging in age from 6 to about 13 years. All students, selected from nine area schools, completed a group-administered test of intelligence, the Kuhlman-Anderson Intelligence Test (N = 521 whites and 423 blacks). A subsample of students also completed two individually administered intelligence tests, the Stanford-Binet (1916 version) and the Arthur Performance Scale (N = 86 whites and 72 blacks).

Mean scores for blacks and whites on each of the three intelligence measures indicate that scores for both groups fell below average (e.g., Stanford-Binet group mean is approximately 100). Subgroup differences indicate that blacks as a group obtained scores 14 to 17 points lower than whites as a group. These results are reported in Table 24.

Because family income levels for blacks included in the study were lower than income levels for whites, Bruce obtained pairs of black and white students matched on economic status and compared mean scores for the two groups (sample sizes not reported). Although mean score differences between blacks and whites are smaller in this sample than in the larger sample, they represent a 9-to 12-point difference (see Table 24). Mean score differences between the two groups were approximately equal for all types of test materials (e.g., general information, novel or new situations, and speed versus power tests).

Evidence provided by Tanser and Bruce indicates that not all differences between blacks and whites can be attributed to educational or economic differences. Tyler (1965) reported that others disagree with this conclusion and argue that the differences may be due to the inadequacy of tests used to measure the intelligence of blacks and that developmental variables other than educational or socioeconomic variables have a depressing effect on the mental growth of black children (Dreger & Miller, 1960; Klineberg, 1963). Empirical evidence to support these arguments has not been collected.

In sum, the question of why these differences appear still remains unanswered. Researchers continue to explore the source or sources of subgroup differences in a variety of ways, such as investigating the influence of developmental, physiological, cultural, and genetic variables. Research investigating each potential source has met with some measure of success as well as with criticism. For example, Levin (1988) describes two lines of research that demonstrated test score gains for black youth. Ramey and his colleagues (1988) met with success by intervening in the preschool years (i.e., provided disadvantaged families with educational and support services). Comer (1980, 1986, 1987) demonstrated that interventions in early elementary grades can help inner-city black youth to raise test scores and even to maintain those gains.

Test Score Interpretation Bias: Differential Validity

Definition of Terms and Procedures. Another potential source of bias involves interpreting test scores. General intelligence and cognitive ability test scores are used to make inferences about potential for success in educational or occupational settings. The soundness of these inferences is determined by the validity coefficient, obtained by regressing criterion performance scores against predictor test scores. The resulting validity coefficient indicates how well one can predict subsequent criterion performance from performance on ability measures. Confidence in prediction is established by statistical methods, assessing whether the validity coefficient is statistically different from zero. If significant, predictor tests may be administered to a new pool of applicants and inferences from test scores can

Table 24

Black and White Mean Score Differences Reported in Bruce Studies

Mean Score Differences for Total Sample of Blacks and Whites[a]

Measure	White		Black	
	N	Mean	N	Mean
Kuhlman-Anderson	521	88	432	72
Stanford-Binet	86	90	72	76
Arthur Performance Test	86	94	72	77

Black and White Students Matched on Economic Level[a,b]

Measure	White Mean	Black Mean
Kuhlman-Anderson	83	73
Stanford-Binet	86	77
Arthur Performance Test	89	77

Summarized from Tyler, 1965.

Note: From "Factors Affecting Intelligence Test Performance of Whites and Negroes in the Rural South" by M. Bruce (1940), Archives of Psychology, 252.

[a]Standard deviations not reported in this study.
[b]Sample sizes not reported.

be used to make selection decisions. Hence, for all applicants, regardless of subgroup membership, test scores are interpreted in the same manner.

Subgroup membership may, however, be an important variable in drawing inferences from test scores. Criterion performance scores are regressed against predictor scores separately for minority and majority groups and the obtained validity coefficients are statistically compared. If subgroup coefficients differ significantly, predictor scores cannot be interpreted in the same manner for the two subgroups. (The appropriate procedure for testing significant differences between two validity coefficients is discussed later in this section.) If inferences are based upon the validity computed for the total group or the majority group when in fact differences exist between majority and minority validity coefficients, then the prediction system is viewed as biased.

Determining the appropriate test for identifying differences between subgroup validity coefficients has often led to controversy. For example, the null hypothesis test, in which each observed subgroup validity coefficient is compared against zero, provides information about the statistical significance of each coefficient; one can conclude that the two differ if the validity coefficient for one subgroup is statistically significant while the coefficient for the other does not differ significantly from zero. A second procedure involves a statistical comparison of subgroup validity coefficients; if the two differ significantly, then test scores cannot be interpreted in the same way for both subgroups.

Confusion about which statistical procedure to use has been addressed by several authors. For example, Humphreys (1973) distinguished between the two procedures described by noting that each answers a different question. The null hypothesis test indicates whether or not predictor measures may be used to draw inferences about subsequent performance for a particular subgroup; these results provide no information about differences between subgroup validity coefficients. A direct comparison between validity coefficients, on the other hand, does provide information about differences between two coefficients.

Humphreys also noted that the two procedures possess different properties. With small sample sizes for one or both subgroups, the likelihood of finding significant differences between subgroup validity coefficients is greater using the null hypothesis test. Directly comparing two validity coefficients requires a sufficient sample size to detect differences between sample correlations from populations that are probably very similar. In most validation studies, minority group sample sizes are typically small; thus the null hypothesis test would indicate that subgroup coefficients differ significantly more often than would the direct comparison method. Because Humphreys (as well as other researchers) believed that majority and minority subgroups represent very similar populations, he concluded that the null hypothesis test leads to erroneous conclusions about subgroup differences, and that direct comparison is the preferred approach because it involves a more rigorous and direct comparison of two validity coefficients.

Boehm (1972) described another statistical procedure used to test for differences between subgroup validity coefficients. This procedure incorporates components of the null hypothesis test and the direct comparisons test. According to Boehm, in this procedure two conditions must exist to demonstrate differences between subgroup validity coefficients: (a) the obtained validity coefficient is significantly different from zero for one group only, and (b) no significant differences exist between the two validity coefficients. Keeping in mind that statistical tests are conducted on sample data to draw inferences about the underlying population, population parameters depicting this phenomenon would be as follows:

$$0 = P_1 = P_2 \neq 0$$

Bartlett, Bobko, and Pine (1977) pointed out that this procedure as a statistical hypothesis about population values is illogical. For example, Part b of Boehm's procedure is satisfied only when the two validity coefficients are exactly equal ($P_1 = P_2$); when Part b is satisfied in the population, Part a cannot be true. Because this phenomenon does not exist in the population, it is illogical to test for it. Bartlett and associates concluded that when this phenomenon is encountered in research, it serves as a warning that insufficient sample information exists to draw inferences about differences between subgroup correlations in the population.

The terms used to describe phenomena observed from statistical test results just described need to be clarified. When results from null hypothesis tests indicate that the validity coefficient for one subgroup is significantly different from zero whereas the validity coefficient for another subgroup does not differ significantly from zero, single-group validity is said to exist. Results from Boehm's procedure also indicate the existence of single-group validity. Differential validity is said to exist when results from a direct comparison test indicate that subgroup validity coefficients differ significantly from one another.

The distinction between single-group validity and differential validity has been best clarified by Boehm (1972). According to her, differential validity exists when "a) there is a significant difference between the correlation obtained for one ethnic group and the correlation of the same device with the same criterion obtained for the other group, and b) the validity coefficients are significantly different from zero for one or both groups" (page 33). Stated in another way: There are four possible outcomes related to observed validity coefficients for majority and minority subgroups: (a) the test is valid for the majority group only, (b) the test is valid for the minority group only, (c) the test is valid for both groups, and (d) the test is not valid for either group. Boehm and others concur that differential validity must be assessed by determining whether or not the validity coefficients differ from one another when at least one validity coefficient is significantly different from zero (outcomes a, b, or c).[19]

Validity Coefficient Differences for Black and White Samples. Evidence related to differential validity has been examined by numerous researchers (e.g., Boehm, 1972, 1977; Katzell & Dyer, 1977; Schmidt, Berner, & Hunter, 1973). In these studies, differential validity for black and white subgroups was examined by accumulating evidence from several studies.

Boehm (1972) analyzed results from 13 studies and found very little evidence of differential validity. Schmidt and associates (1973) concluded from a review of 19 studies that differential validity, when it appears, occurs by chance and is due to defects in statistical procedures used. Thus, according to these authors, differential validity is a pseudo-problem.

Boehm (1977) analyzed data from 31 studies and concluded that differential validity is rare. She went on to state that an unequivocal test of the issue has not been conducted because of low statistical power and other deficiencies in accumulated studies. Katzell and Dyer (1977) determined from their review of 31 studies that differential validity is not a pseudo-problem and that researchers should continue to check on the phenomenon.

Linn (1978b) summarized the evidence from these reviews of studies to establish some closure on the issue. He concluded that the evidence indicates that differential validity is rare and that, in general, when it occurs, differences between validity coefficients for blacks and whites are small.

Validity Coefficient Differences for Male and Female Samples. Results from research examining differences between validity coefficients computed for male and female samples are less conclusive. In a study to investigate these differences, Schmitt, Mellon, and Bylenga (1978) accumulated more than 6,200 male-female validity coefficient pairs. Analysis of these data included comparing mean validities for males and females (a) computed across all pairs; (b) computed separately for different types of predictor measures (e.g., cognitive ability tests, personality measures, biographical inventories); and (c) computed by criterion measure--educational and occupational.

Results indicate that across all coefficient pairs, validities for female samples exceeded validities for male samples by .04 correlational unit (SD = .20). Median validity estimates computed by predictor type indicate that values for males and females differ most on cognitive ability measures. Specifically, validity coefficients are higher for females than males on measures of verbal ability (mean difference = .04 and SD = .20, computed across 1,950 validity pairs), abstract reasoning (mean difference = .05 and

---

19 Differential validity is assessed by transforming the validity coefficients to Fisher's Z-values and then using the following formula to calculate a critical value:

$$\text{Critical ratio} = Z_1 - Z_2 / \sqrt{1 / N_1 - 3 = 1 / N_2 - 3}$$

SD = .13, computed across 1,839 validity pairs). The smallest differences appeared on personality measures; validities for males are only slightly higher than validities for females (mean difference = .003 and SD = .13, computed across 80 validity pairs). Mean differences between validity coefficients computed for male and female samples, using an academic criterion indicate that females are slightly more predictable than males (mean difference = .04 and SD = .20, computed across 6,053 validity coefficients). Males, on the other hand, appear slightly more predictable when measures are validated against employment criteria (mean difference = .04 and SD = .22, computed across 135 coefficients).

Overall, the authors concluded that females appear to be slightly more predictable than males. This difference reflects only .04 correlational unit, thus it may reflect only a trivial difference when viewed from a practical standpoint. Firm conclusions cannot be drawn from these data because, as Schmitt and associates note, many of the studies from which the validity coefficients were obtained included small sample sizes. Thus, statistical power to detect true differences between male and female validity coefficients is low. Although differences between validity coefficients for male samples versus female samples in this review of studies are small, it may be informative for researchers to continue investigating differences between male and female validity coefficients, especially when large sample sizes are available.

Test Score Interpretation Bias: Differential Prediction

Although Federal guidelines for employment selection practices require researchers to compare subgroup validity coefficients (e.g., for blacks and whites; for males and females), the lack of differential validity fails to provide sufficient evidence to conclude that test interpretation is unbiased (Bobko & Bartlett, 1978; Humphreys, 1973). In other words, concern about bias in test score interpretation can best be answered by comparing the entire prediction system for different groups. This source of bias was previously referred to as differential prediction.

Demonstration of bias due to differential prediction involves generating regression equations separately for each subgroup and then comparing subgroups' regression slopes, regression intercepts, and standard errors of estimate about the regression line. If significant subgroup differences appear on <u>one or more</u> of these components--slopes, intercept, and standard error of estimate--then different prediction systems are required to interpret test scores for each subgroup. If a common regression equation is used in this situation, bias is said to occur.

Several studies have been conducted to assess the frequency with which differential prediction occurs. Bobko and Bartlett compared slope and intercept differences for more than 1,190 majority and minority subgroup regression equations. They reported that 68 (5.2 % of the 1,190 comparisons) exhibited significant differences in slope values and 214 (18 %) exhibited significant differences in intercept values. One may be tempted to conclude from these data that different regression equations would be required for majority and minorities in about 282 (23.2 %) of the 1,190 values computed.

Because many of the equations included in these analyses were pulled from the same studies, the equations do not represent independent data sets. Therefore, Bobko and Bartlett concluded that the actual frequency with which differential prediction occurs cannot be determined from these data.

Jensen (1980) also addressed this issue by reanalyzing data provided by Ruch (1972), who compared differential prediction equations generated for blacks and whites reported in 20 studies. In the reanalysis, Jensen tallied the number of times: (a) the standard error of estimate, the slope, or the intercept were non-significantly different ($p > .05$), (b) one or more of three components was significantly larger for whites than blacks ($p < .05$), and (c) one or more of the components was significantly larger for blacks than for whites ($p < .05$). Using these data, Jensen determined whether significantly different slopes, intercepts, and standard errors of estimate, when they occur, consistently favor one group over another or favor both groups with equal frequency. If the direction of bias (or significant differences between subgroup regression components) is random, then no significant differences between the frequencies of white greater than black or black greater than white will appear.

Results (Table 25) from these analyses indicate that for slopes and standard errors of estimate, differences between frequencies of occurrence are non-significant (i.e., W > B = B > W). Thus, across studies of bias in standard errors of estimate and slopes, there is no evidence to suggest that selection decisions will consistently favor one group over another.

Table 25

Summary of Black and White Differences in Regression Parameters in 20 Independent Studies

Regression Parameter	Total	Non-significant	Significant ($p < .05$) W > B	Significant ($p < .05$) B > W	X^2
Standard Error of Estimate	20	12	5	3	.50 (NS)
Slope	20	9	7	4	.82 (NS)
Intercept	20	8	11	1	8.33 ($p < .01$)

Note: From Bias in Mental Testing by A. R. Jensen (1980), New York: The Free Press. (Copyright 1980 by The Free Press.) Reprinted by permission.

This same conclusion does not apply to intercept differences. There _is_ significant and consistent bias for intercepts, with intercepts for whites more frequently higher than intercepts for blacks. According to Jensen, if a regression equation generated on a white sample is used to predict criterion performance for blacks, more often than not it overpredicts blacks' average performance.

In general, then, when regression equation components for majority and minority subgroups are compared, significant differences appear more frequently between subgroup intercepts than between subgroup slopes or standard errors of estimate. According to Dunnette and Borman (1979), significant intercept differences are due to subgroup mean differences between predictors, criteria, or both.

In general, evidence for subgroup differences on cognitive ability measures indicates that minority mean scores will be from 0.50 to 1.00 standard deviation units below that of the majority mean. Although differential validity seldom appears between minority and majority group validity coefficients, differences in regression equations do appear, most often because of intercept differences.

Most frequently, significant intercept differences appear, indicating that bias in test interpretation may occur if a common regression equation is used. A similar but not identical situation occurs when comparing male and female test scores and prediction equations. That is, cognitive ability mean test scores may differ very little or may differ in some cases up to one standard deviation depending upon the cognitive ability assessed. Evidence available at this time, however, indicates that validity coefficients for these two subgroups differ very little.

Test Fairness

Thus far, we have summarized the evidence regarding bias in interpreting test scores. Data reviewed indicate that bias may occur because of differences between subgroup slopes and intercepts. If bias exists, one must then decide how to utilize test information to ensure fairness in selection decisions. In other words, a primary goal in drawing inferences from test scores (whether or not test bias has been demonstrated) is to ensure that members of all groups have an equal opportunity for selection, given equal ability to perform well or to succeed in educational or occupational settings. Test fairness issues attempt to address this goal.

Models of Test Fairness. Numerous researchers have developed procedures or models to specify what constitutes test fairness. The models vary with respect to social, philosophical, and legal considerations as well as statistical procedures. The models also place different emphases on making correct decisions versus avoiding incorrect decisions and differ with respect to criterion performance outcomes. Thus, one way to compare and contrast test fairness models is to examine outcomes such as hits (true positives and true negatives), misses (false positives and false negatives), and average criterion performance of the selected group.

Below we describe five models depicting test fairness guidelines for interpreting test scores to ensure equal opportunity. We then present a hypothetical situation in which all the models are used to address test fairness. Outcomes such as hits and misses and average criterion performance are described for each model.

(1) Cleary (1968): Regression Model

According to the Cleary model, inferences drawn from test scores are biased if the use of a common regression equation for all subgroups results in consistent non-zero errors of prediction for members of one subgroup. Hence, a test is biased if the criterion score predicted from a common regression line is consistently too high or too low for members of one subgroup. If consistent non-zero errors appear with a common regression equation, the recommended procedure would be to utilize separate regression equations for each subgroup and to select those with the highest predicted criterion scores.

(2) Einhorn and Bass (1971): Equal Risk Model

This model indicates that a test is fair if the risk or probability for success is equal in both groups. Thus, for each group, predictor cut-off points are set above which applicants have a specific chance for success. To establish the predictor cut-off for each group, one first establishes the maximum probability of a selection error as defined by the false positive rate (risk) one is willing to accept, given the predicted criterion score (i.e., y).

For example, the risk or probability of an error may be set at 20 percent for each group. From a normal probability curve the z_p value would be set at -.53. The z_p values for members of each group are computed using the following formula:

$$z_p = \frac{(y^* - y)}{SEy}$$

where: z_p = deviate from the normal curve

y^* = criterion of success-failure threshold

y = applicant's predicted score on the criterion

SEy = standard error of estimate of y in the applicant's group

From this formula, applicants obtaining scores (z_p values) greater than -.53 are rejected while those obtaining scores (z_p values) lower than -.53 are accepted.

(3) Thorndike (1971): Constant Ratio Model

According to the constant ratio model, a selection measure is fair if the ratio of the probability of success on the criterion for two groups is equal to the ratio in which the groups are selected. For example, if data from the criterion measure indicate that 60 percent of Group A perform successfully and 40 percent of Group B perform successfully, then the selection system should reflect the same selection ratio. In this case, 60 percent of Group A and 40 percent of Group B are selected.

(4) Cole (1973)/Darlington (1971): Subjective Regression/Conditional Probability Model

According to Darlington, if X represents the predictor measure, Y the criterion measure and C the cultural variable (scored 0 for minority and 1 for majority groups), the test is fair if:

$$r_{xc.y} = 0$$

Thus, the partial correlation between test scores and cultural group membership with criterion scores parceled out should be equal to zero. If not, this indicates that greater differences between cultural groups appear on the predictor measure than would be predicted by the criterion. Hence, if the mean criterion scores for the two groups are equal or very similar and the mean predictor scores differ significantly (with the majority group scoring lower), then $r_{xc.y} \neq 0$. In this situation, the probability of selection, given a criterion level pass point, would be lower for the majority group than the minority group because of the mean difference between the groups on the predictor scores.

To ensure fairness, predictor scores for persons in the lower scoring group are adjusted to make certain that minority and majority group members with the same criterion scores (indicating probability of success) have the same predictor scores (indicating probability of selection). Thus, the probability of selection, given a specified level of criterion performance, is equal for all persons regardless of group membership.

(5) Quota Model

To follow the quota model, the proportion of minorities selected in educational or occupational settings should reflect the same proportion as minorities in the population. Test users may define the population in one of several ways, such as population rates, regional rates, or the proportion of minorities in the applicant population. The quota system may then be implemented by rank ordering applicants according to test scores within subgroups. The number selected from majority and minority subgroups is a function of total numbers of positions to be filled and subgroup representation in the defined population.

Comparison of the Models. To demonstrate the varying effects of the models, Dunnette and Borman (1979) described a hypothetical situation in which a criterion-related validity study has been conducted for 200 male and female telephone operators. Validity coefficients computed separately for males and females are of moderate size and do not differ significantly from one another. The mean predictor test score for males is one standard deviation below that of the female predictor mean, and the criterion mean for males is one-half standard deviation below that of the criterion mean for females. Table 26 provides the selection results for each of the five models discussed above including: (a) the procedures used to interpret test scores for each subgroup, (b) a definition of fairness or lack of fairness with respect to a specific model, and (c) the proportion of members of each subgroup selected and resulting job performance levels.

As noted in the Results section of the table, the models give different weights to the benefits and costs associated with different selection errors. According to Dunnette and Borman, the regression model maximizes average criterion performance of selectees and minimizes the risk of job failure while denying employment opportunities disproportionately to potentially successful persons from different subgroups. The quota model, on the other hand, provides employment opportunity equally to members of all subgroups but results in lower average criterion performance, disproportionate subgroup risk of failure, and disproportionate subgroup rejection of potentially successful persons.

This hypothetical example makes it clear that outcomes vary according to the fairness model selected. Decisions about which model best represents test fairness require test users to weigh and evaluate each outcome. For example, test users emphasizing productivity outcomes or high average criterion performance would most likely use the Cleary model. On the other hand, test users placing more emphasis on outcomes beneficial to individuals or particular groups may opt for the Quota model. Selecting the appropriate fairness model, then, requires users to identify and evaluate outcomes, both organizational and individual, and to consider social, political, philosophical, and legal issues.

While it is not within the realm of this report to provide a definitive statement about the "best" fairness model, we do provide a recommendation for practical consideration. Results from a study by Hunter, Schmidt, and Rauchenberger (1977) indicate that the Cleary model yields the highest average criterion performance when compared with the Thorndike, Darlington, and Quota models. The Quota model allows for the highest minority selection rates when compared with the other models. Thorndike's model, however, represents a compromise to the Cleary and Quota models by yielding average criterion performance values nearly as high as those observed using Cleary's model while, at the same time, increasing minority selection rates. Thus, the Thorndike model, compared to Cleary's model, results in a selected group with high average criterion performance while at the same time increasing minority representation in educational or occupational settings.

In a subsequent review of test fairness models, Schmidt (1988) argued that most models are actually disguised quota systems. Because quotas are generally lower for blacks than whites, adverse impact is reduced but not eliminated. Moreover, in recent years, all models with the exception of one have fallen into disfavor. That is, APA Standards for Educational and Psychological Testing (1985) refer to only the regression model of test fairness.

Table 26

Models of Test Fairness: Description of Test Use, Test Fairness, and Results of the Selection Process

Model	Test Use	Fairness	Results
Cleary or Regression Model	Use separate regression equations for males and females. Persons with highest predicted performance are selected.	Test is unfair only if consistent non-zero errors of predeiction are made for members of one subgroup.	Average criterion performance is high. Proportion of males selected equals 2:8. 'Proportion of false positives is equal for each group Proportion of false negatives is higher for males than females.
Einhorn and Bass or Equal Risk Model	Use separate regression equations to determine expectancy charts for males and females. Cutting scores are set to yield the same levels of of success or failure (risk) for both groups.	Test is unfair if expected success and failure rates differ according to subgroup membership.	Results are essentially the same as the Regression Model.
Thorndike or Constant Ratio Model	Points equivalent to one-half S.D. are added to each male applicant's test score. Revised scores are interpreted using a common regression equation.	Criterion scores define the true potential for each subgroup, thus the test score distribution is modified to be the same as the criterion.	Average criterion performance is somewhat lower for the selected group than that for persons selected using the regression model. Proportion of males selected equals 4 out of 10. Proportion of false positives is much higher for males. Proportion of false negatives is equal for both males and females.
Cole/Darlington Model	Separate regression equations are generated for each group. The predictor score distribution is *modified to assure that males and* females with the same criterion score have the same *predictor score*. Scores are interpreted according to a common regression equation.	Test is defined as unfair if it excludes from selection unequal proportions of potentially suc*cessful persons from different* subgroups.	Average performance of selected group is lower than that of group selected by the Regression Model or *the Constant Ratio Model*: *Proportion* of males selected equals 4 out of 10. Proportion of false negatives is equal for both males and females.
Quota Model	Separate regression equations are computed *for males and females. The* same selection ratio is used for both because the proportion of male applicants is equal to the proportion *of female applicants*.	Test is defined as unfair if a *different proportion of persons* is selected from different subgroups.	Average criteria performance of all *selected persons is about the same as* the Cole/Darlington Model. Proportion of males equals the proportion of females selected. Proportion of false positives is much higher for males. Proportion of false negatives is lower for males than females.

Note: From "Personnel Selection and Classification Systems," by M. D. Dunnette and W. C. Borman (1979), Annual Reviews of Psychology, 38. (Copyright 1979 by Annual Reviews, Inc., Palo Alto, CA). Summarized by permission.

Summary

In this part, we examined several sources of test bias and models designed to ensure fairness in test score interpretation. A review of the cited literature concerning bias in test content, suggests that cultural bias theory, as set forth, does not account for mean test score differences between minority and majority subgroups. Operational definitions that explicitly describe the sources of cultural bias are needed.

We also examined test bias in statistical terms. Differential validity exists when the validity for one subgroup differs at a statistically significant level, from the validity computed for another subgroup. A summary of the literature investigating the frequency with which minority and majority validity coefficients differ suggests that differential validity is rare, and when it appears differences are small. Differences between validity coefficients computed for males and females are also small. Nevertheless, it is instructive to examine differential validity when sample sizes permit.

In addition to examining differential validity, researchers are also advised to examine differential prediction. This involves comparing slopes, intercepts, and standard errors of estimates for minority and majority subgroups. The literature suggests that differences between regression equations computed for blacks and whites appear most frequently for the intercept. Thus, bias in test score interpretation may occur if a common regression equation is used.

Finally, five models which have been developed to specify test fairness in a selection situation were reviewed and compared in terms of outcomes, such as correct decisions and average criterion performance. The Federal guidelines which are described in the last part of this section, indicate that test users must examine test fairness.

UNIFORM GUIDELINES AND LEGAL IMPLICATIONS: IMPACT ON COGNITIVE ABILITY MEASUREMENT

Concerns about bias in tests have led to a vast amount of research centering around use of tests for selection purposes. In recent years, the Federal government has demonstrated concern about the use of tests to make selection decisions. Much of this concern relates to a practice of discriminating against various minority groups or "protected classes" in selection decisions. Discriminatory practices such as this were determined to be illegal with the passage of the Civil Rights Act in 1964.

We review the features of the Civil Rights Act related to the use of tests for employment purposes, the original and revised uniform guidelines established to develop and implement selection systems, and major court decisions based upon interpretation of the guidelines which further serve to guide selection system development.

The Civil Rights Act and Title VII

The Civil Rights Act of 1964 established that discrimination in various sectors of our society is forbidden; we focus here on one part of the Act, Title VII. This Title, which deals specifically with discrimination in employment, states:

> It shall be an unlawful employment practice for an employer (1) to fail or to refuse to hire or to discharge any individual or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment because of said individual's race, color, religion, sex, or national origin; or (2) to limit, segregate or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee because of such individual's race, color, religion, sex, or national origin.

Title VII called for the establishment of the Equal Employment Opportunity Commission (EEOC). It was this agency that first prepared and published Guidelines on Employee Selection Procedures (EEOC, 1966) and later published Guidelines on Employee Selection Procedures (EEOC, 1970). The EEOC was charged with enforcing Title VII, including monitoring selection programs for all employers with 15 or more employees, labor unions engaged in "industry affecting commerce," employment agencies that serve the above industries, state and government agencies, and educational institutions. An amendment to the Civil Rights Act in 1972 provided for the establishment of the Equal Employment Opportunity Coordinating Council (EEOCC). This council included the Secretary of Labor, the Attorney General, the Chairman of the Civil Service Commission, and the Chairman of the Civil Rights Commission. The EEOCC was charged with establishing guidelines for the four agencies represented by the council--the Department of Labor, the Department of Justice, the Civil Service Commission, and the EEOC. The Uniform Guidelines on Employee Selection Procedures were published jointly by the four agencies somewhat later (EEOC, 1978). Below we review the guidelines established in 1970 and compare them with the more recent Uniform Guidelines.

Guidelines on Employee Selection Procedures

In the first set of guidelines, the term, test, was defined as "any paper-and-pencil measure or performance measure used as the basis for an employment decision," which includes eligibility for hire, transfer, promotion, membership, training, referral, or retention. According to the 1970 guidelines, a test includes but is not limited to measures of general intelligence, mental ability and learning ability, specific intellectual (cognitive) abilities, dexterity and coordination, occupational interests, attitudes, personality, and temperaments. Thus, it appears that any type of instrument or tool designed to assess human characteristics for purposes of employment selection is considered a test.

Discrimination is defined as the use of any test that adversely affects employment opportunities of classes protected by Title VII. Test use in this case would be considered unlawful unless (a) the test has been validated and evidences a high degree of utility, and (b) the person giving or acting upon the results of the particular test can demonstrate that alternative suitable hiring, transfer, or promotion procedures are not available for use (EEOC, 1970). Discrimination is also demonstrated when minority candidates are rejected at a higher rate than non-minority candidates. When it is technically feasible, or when sufficient sample sizes are available, a test should be validated separately for each minority group. Differential rejection rates can be justified by demonstrating relevance to performance on the job.

Validity for a particular selection test "must be based on studies employing generally accepted procedures for determining criterion-related validity" (EEOC, 1970, p. 12333). These earlier guidelines recognized that in situations where it is not technically feasible to conduct a criterion-related validity study (e.g., due to small sample sizes), a content or construct validity approach may be used.

Other minimally acceptable standards of criterion-related validity studies outlined by the 1970 guidelines include:

> The study sample must be representative of the normal or typical population of candidates for the job in question.
>
> Tests must be administered and scored following standardized procedures with proper safeguards to ensure test security.
>
> The work behaviors or other criteria of employee adequacy must be fully described.
>
> In view of possible bias inherent in subjective performance evaluations (e.g., supervisory ratings), these measures must be carefully developed and the resulting data examined for evidence of bias.
>
> Validity coefficients and other data should be computed separately for minority and non-minority groups whenever technically feasible.

The Uniform Guidelines

The most recent guidelines, published and endorsed by the Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, differ to some degree from the first set of guidelines. For example, discrimination is defined with more clarity by quantifying the term adverse impact. According to this guidance, adverse impact occurs when the selection rate for any race, sex, or ethnic subgroup is less than four-fifths (80 %) of the rate for the group with the highest rate of selection. According to this definition, if the selection rate for the majority group is 50 percent, a selection rate for the minority group should be at least 40 percent to demonstrate a lack of adverse impact. If the minority selection rate were to be less than 40 percent (e.g., 20 %), it would be regarded by Federal enforcement agencies as evidence of adverse impact.

Another standard established by the uniform guidelines involves the "bottom line" approach to assessing adverse impact in the selection system. If a selection system contains several stages of testing in which candidates are accepted or rejected at each stage, the employer need only compare the selection rates for the total selection system. If this comparison provides evidence for adverse impact, then the employer is required to compare the selection rates at each stage (or for each component) and remedy the situation where evidence for adverse impact exists.

The guidelines also stipulate that when two or more selection procedures are available and both are equally valid, the employer should select the procedure having the "lesser adverse impact." While conducting a validity study, then, one should investigate suitable alternative selection procedures.

The uniform guidelines describe procedures for conducting criterion-related validity studies in more detail than the earlier guidelines. The 1978 guidelines emphasize the need for job analysis to determine the relevant or critical work behaviors required in the target job. This information is then used to develop criterion measures that represent the important components of the job. Criteria developed without a full job analysis may be used if the employer can demonstrate their importance to the particular employment context. Criteria include but are not limited to production rates, error rates, tardiness, absenteeism, and length of service. Unlike the first set of guidelines, the most recent guidelines indicate that content or construct validation strategies will be viewed favorably by the agency. Detailed procedures for using each of these validation strategies are provided in the current guidelines.

The guidelines also call for examining unfairness in validation studies, defining unfairness as the situation in which "members of one race, sex, or ethnic group characteristically obtain lower scores on the selection procedure than members of another group and the differences are not reflected in differences in a measure of job performance. Use of the selection procedure may unfairly deny opportunities to members of the group that obtains the lower scores." Arvey (1979) pointed out that in comparison with the earlier guidelines "this definition reflects both a more sophisticated treatment of the fairness issue and avoids any major focus on differential validity" (p. 79).

The guidelines recognize the feasibility and practicality of utilizing less common approaches to test validation. For example, employers are encouraged to participate in "cooperative" or consortium validity research. Validity evidence obtained using this approach is evaluated by the validity for a target job as a whole and not by the validity specific to each participating organization. In addition, the standards spell out in detail the requirements for using "borrowed" studies to generalize validity results to other jobs.

Finally, employers are encouraged to design and implement affirmative action programs to remedy past discriminatory hiring practices. Although such programs are voluntary and no strict standards exist for designing such a program, the guidelines encourage employers to consider several components of the selection system, including recruitment programs, work or job requirements, selection instruments, and career advancement opportunities. The guidelines specify that an affirmative action program does not require one to employ unqualified persons nor does it require selection of persons based on race, sex, religion, or national origin.

Key Judicial Decisions

As has been described in the recent Uniform Guidelines, standards for conducting validation studies and implementing selection systems are spelled out in much more detail and have been expanded from earlier guidelines. Detailed information about procedures has been added to clarify standards. Further, some modifications are in response to judicial interpretations of the earlier guidelines. As early as 1963, cases involving discrimination in the employment setting were reviewed by the courts. In each case, the presiding judge must determine how closely to follow the Uniform Guidelines. Table 27 lists some of the notable court cases along with the important decisions provided in each.

The first case appearing in this table, Myart v. Motorola (1964), established a precedent for hearing employment cases in the court system. The remaining cases cited provide details about judicial interpretation of the law, changing trends in the court's adherence to the EEOC guidelines, and the high level of sophistication involved in judicial evaluation of validation studies. The decisions are summarized briefly as follows:

- Employers may continue to use professionally developed tests but these tests must be validated by the employer (Hicks v. Crown Zellerbach Corporation, 1970; Griggs v. Duke Power Company, 1971). Along similar lines, tests used for employment purposes must be developed by professionals who have training in psychological testing.

- Courts consider the comprehensiveness of a job analysis and require documentation to support the comprehensiveness. When criterion performance appraisal forms are being developed, the job behaviors rated must be specified, as opposed to simply rating overall performance (Albemarle v. Moody, 1975). In addition, training scores represent an acceptable performance criterion against which test scores may be validated (Washington v. Davis, 1976).

- Tests intended for selection at the entry-job level must be validated at that level. Further, the sample included in the validation study must be representative of candidates applying for the target job (e.g., similar age, race, and sex composition) (Albemarle v. Moody, 1975).

- Courts demonstrate a sophisticated understanding of test validation principles and tools. For example, the court has considered

Table 27

Decisions From Significant Court Cases

Case	Decisions or Outcomes
1. Myart v. Motorola (1964)	Established a precedent for hearing employment cases in the court system.
2. Hicks v. Crown-Zellerbach Corporation (1970)	Professionally developed tests may be used to assess applicants' qualifications but the user must demonstrate the job-relatedness or validity of the measures.
3. Griggs v. Duke Power (1971)	Discriminatory <u>intent</u> is not the issue in Title VII cases, instead the <u>consequences</u> of employment practices are the focus. Employers must demonstrate business necessity for using measures and demonstrate that hiring decisions are based on job-related factors. Great deference to EEOC Guidelines is acknowledged.
4. Diaz v. Pan American Airways (1971)	Bona fide occupational qualification (BFOQ) exceptions are narrowly interpreted. Sex-role stereotypes or preferences by employers, clients, or customers do not warrant BFOQ exceptions. Business <u>necessity</u>, not business <u>convenience</u>, is the issue.
5. U.S. v. Georgia Power Company (1973)	EEOC Guidelines provide a framework for evaluating validation studies and the validity study under review suffered from several technical flaws, which included use of an inappropriate study sample, and no investigation of differential validity although it was technically feasible.

(Continued)

Table 27 (Continued)

Decisions From Significant Court Cases

Case	Decisions or Outcomes
6. Albemarle v. Moody (1975)	Tests validated for a single job are not valid for jobs at other levels. Thus, tests validated on incumbents in middle- or top-level jobs are not necessarily valid for entry-level applicants.
	Criterion performance measures (e.g., performance evaluations) must include clear definitions of the behavior to be rated and guidelines for providing the ratings.
	Tests must be validated on a sample representative of the applicant sample.
7. Washington v. Davis (1976)	Plaintiffs filing complaints under the 5th Amendment must demonstrate intent to discriminate; demonstration of adverse impact is insufficient.
	Training performance scores may serve as criterion performance measures to demonstrate the job-relatedness of a test.
8. Bakke v. University of California at Davis (1978)	Equal protection (14th Amendment) cannot be limited only to protected groups.
	Establishing a system to insure that economically disadvantaged individuals are given the opportunity to higher education is worthwhile. Thus, race may be used as a factor in determining admissions. A strict quota system is, however, inappropriate.
9. United Steelworkers of America v. Weber (1979)	Voluntary affirmative action programs that utilize quotas to eliminate racial imbalances are permissible.
10. Connecticut v. Teal (1982)	Even if the "bottom line" approach indicates no adverse impact occurs, adverse impact in one component of the selection system constitutes a discriminatory practice.

statistical procedures used to demonstrate criterion-related validity and assessed the technical feasibility of investigating differential validity (U.S. v. Georgia Power Company, 1973).

- Bona fide occupational qualifications (BFOQ) or discriminatory practices on the basis of race, sex, religion, or national origin for business reasons are viewed very narrowly. Thus, preferences by employers, clients, customers, do not warrant BFOQ exceptions. Employers must demonstrate business necessity[20], not business convenience, for BFOQ exceptions (Diaz v. Pan American Airways, 1971).

- The notion of discriminatory intent need not be demonstrated in Title VII cases, only the consequences of discrimination or evidence for adverse impact are required (Griggs v. Duke Power, 1971). Complaints filed under other amendments or acts other than Title VII may be required to demonstrate intent to discriminate (Washington v. Davis, 1976).

- Affirmative action programs are encouraged by the courts; thus, voluntary programs that utilize quotas to eliminate racial imbalances are permissible (Steelworkers v. Weber, 1979). Another court determined that strict quota systems which result in reverse discrimination may be viewed as inappropriate (Bakke v. Regents of the University of California at Davis, 1978). Hence, the status of quota systems in affirmative action programs is unclear.

- Evidence for adverse impact may be obtained by comparing selection rates for minority and non-minority groups at each stage of testing. Thus, the absence of adverse impact at the "bottom line" does not necessarily indicate lack of discrimination (Connecticut v. Teal, 1982). (Note that this decision differs from the Uniform Guidelines "bottom-line" standard.)

Finally, one major trend observed in the court cases concerns the attention given to the guidelines. In early decisions, courts acknowledged great deference to the guidelines (Griggs v. Duke Power, 1971; U.S. v. Georgia Power, 1973). Results from subsequent court cases, however, indicate that judges often view them simply as guidelines that allow for interpretation.

The Uniform Guidelines and subsequent major court decisions offer implications for test development and implementation in employment settings. First, thorough job analyses are required to identify the critical job performance requirements. Results from the job analysis should be used to develop criterion measures that explicitly define the important work behaviors of the target job(s). Second, professionals familiar with psychological testing principles are required to develop selection tests. Third, the

---

[20]According to Cascio (1978), "for discriminatory practice to be allowed as a 'business necessity' that practice must be essential to the safe and efficient operation of the organization. Furthermore, no alternative policies or practices must be available which would be less discriminatory" (p. 25).

validation study must include: (a) standardized procedures for administering and scoring the test, (b) samples that are representative of the applicant population, and (c) examination of the criterion measure (especially subjective ratings) for bias. Fourth, data analysis should include computation of results (e.g., means, standard deviations, and validity coefficients) separately for different subgroups, if sample sizes permit. Finally, test fairness should also be examined.

The preceding description of the recommended procedures to validate and implement selection systems provides only a brief indication of the standards established by the Equal Employment Opportunity Coordinating Council. Specific details and requirements may be obtained from the Federal Register (Friday, August 25, 1978, Part IV). Finally, although the emphasis of this section is on measuring cognitive ability, the procedures outlined apply to the development, validation, and implementation of all types of instruments used to make selection decisions.

Summary

The original Selection Guidelines describe the minimally acceptable standards for constructing and implementing a selection system. The Uniform Guidelines define discrimination more clearly by quantifying the term "adverse impact," outline the specific requirements of validity studies, and recognize the feasibility of less common approaches to test validation, such as consortium validity research.

Court decisions involving EEOC cases indicate how the laws and Guidelines are interpreted. Several key court cases and their implications for validation research were discussed. Decisions for future EEOC cases will provide information about the status of the Guidelines and programs such as affirmative action programs, designed to compensate for previous discriminatory practices.

SECTION SUMMARY AND CONCLUSIONS

This section continues our emphasis on the need for conserving human talents in the work place. The issue in this case involves discriminatory practices that may prevent some capable persons from qualifying for educational or occupational opportunities because of characteristics unrelated to job or educational requirements, such as minority group membership. Specifically, ability tests used to screen applicants were viewed as possible sources of discrimination. Researchers and test developers have been concerned with the issue of discrimination since the onset of intelligence and ability testing.

We examined differences in subgroup mean scores on measures of general intelligence and specific cognitive abilities. Methodological considerations for comparing subgroups on cognitive ability measures were presented. Overall, significant differences were found between male and female mean scores and between majority and minority racial/ethnic mean scores. This is not to suggest, however, that we can predict a single individual's test score from subgroup membership status.

The literature suggests that males and females differ little in general intelligence, but differ to greater extent on specific cognitive abilities. For example, males and females differ on spatial abilities, but this difference varies by item type or task. On measures requiring three-dimensional rotation, males outscore females by about one standard deviation, while measures that require only visualization result in a much smaller difference between the two groups.

Mean test score differences between majority and minority subgroups yield similar patterns for measures of general intelligence and of specific cognitive abilities. On the average, Orientals' mean scores are very similar to or slightly below the white mean score; American Indians and Hispanics score about one-half standard deviation below whites; and blacks score about four-fifths of a standard deviation to over one standard deviation below whites. Measures yielding smaller subgroup differences will receive priority consideration when selecting constructs and developing tests to supplement the ASVAB.

In this section we also examined the Federal guidelines for ensuring non-discriminatory practices in selection system implementation. Although these regulations are not limited to the area of cognitive abilities alone, we described them here because such guidelines will be used to evaluate all new measures designed to supplement the current military selection system.

A primary objective of the literature review is to identify constructs that add unique predictive variance to the current Army selection and classification system. Thus, validity information obtained from past research efforts will also be used to evaluate constructs considered for inclusion in an experimental battery. These data are summarized in the next section.

SECTION V

SUMMARY OF VALIDITY DATA

In this section, we describe the steps involved in summarizing the validity coefficients gleaned from the literature. Summary validity tables are then presented and discussed with respect to implications for the present research project.

The literature search described in the preface to this report resulted in the identification of approximately 4,410 potentially relevant citations. All citation abstracts were screened and evaluated for relevance, a process that identified an initial group of 880 documents for possible review. Closer inspection indicated that approximately 420 were of greatest interest, and these were reviewed.

Approximately 400 Article review forms summarizing each article, technical report, or test manual were completed. Data reported for cognitive predictor measures were recorded on a separate Predictor review form (e.g., test description, reliability estimate, validity coefficient, correlations with other measures). One Predictor form was completed for each predictor described in a validity study, and well over 600 Predictor review forms were completed. Data recorded on these forms provided the validity information summarized in the tables that follow.

PREPARATION OF THE VALIDITY SUMMARY TABLES

Before examining the validity tables, we describe the decision rules used to identify information for the tables and the procedures used to organize the information.

Decision Rules for Including Studies

Predictor Type. One of the first decisions involved determining the type of predictor to include in the tables. For purposes of comparing results across studies, we chose to include results for traditional paper-and- pencil tests only. Thus, tests requiring special apparatus such as tape recorders, headphones, computer equipment, or slide projectors were excluded from this summary.

Tests designed to assess very specific abilities, such as achievement in physics and chemistry courses, or potential to learn a foreign language, were also excluded from this summary. The purpose of the literature review was to identify predictor measures that might be useful for a wide variety of jobs. Tests designed for specific purposes would be applicable for only a very few MOS and were, therefore, omitted.

Predictor measures included in the summary tables, then, represent traditional paper-and-pencil measures of cognitive/perceptual abilities. Because the current military selection and classification battery, the Armed

Services Vocational Aptitude Battery, contains measures of technical knowledge (e.g., Electronics Information, Auto/Shop Knowledge), data for these types of measures were also summarized.

Sample Composition. In the literature search process, we examined a wide variety of studies that described predictor development and/or validation results from a variety of subject populations. Since our purpose was to learn as much as possible about the different cognitive ability measures used for prediction purposes, the subject population was not a major determining factor in identifying studies for review.

In summarizing the validity data, however, our objective was to identify measures that might be used to predict training and job performance outcomes for the Army applicant population. This group includes, for the most part, persons between the ages 18 to 23, who have graduated from high school. Thus, in screening the reviewed data for inclusion in the summary tables, the nature of the subject population was a critical factor. Studies that included young children or persons in college were excluded from the summary validity tables.

Studies involving young children were excluded because measures developed for these samples usually did not reflect the types of ability measures suitable for the Army population (e.g., tests were too easy). Studies that involved college students were excluded for several reasons: (a) the mean age of these samples often exceeded the age range for the Army sample; (b) measures developed for these samples were geared toward higher ability levels (e.g., too difficult for high school populations) or were written to assess very narrow abilities (e.g., knowledge of physics); and (c) validity coefficients for relevant measures administered to college samples suffered from restriction in range.

Another factor was the time period at which the data were collected. We reasoned that older studies, such as those conducted before 1960 or so, often used restricted subject populations; these samples do not necessarily reflect the minority or gender composition of the present work force or military population. Therefore, we focused on the more recent studies reporting validity coefficients for cognitive ability measures.

In using this decision rule, however, we allowed some flexibility in selecting studies. For example, Egbert and associates (1958) conducted a study to examine soldiers' performance under combat conditions in Korea. Because this study represents a comprehensive effort to identify predictors of combat effectiveness and employed criterion measures obtained under actual combat conditions, we elected to include these data in the summary tables. For the most part, other studies prior to 1960 that are included in the summary tables were conducted in a military setting and provide information about jobs or MOS that otherwise would not have been represented in the tables.

In sum, we screened the reviewed literature to ensure that the reported validity coefficients are representative of the validities one might obtain with the target population of Army recruits with respect to age, gender, and minority status. Appendix B contains the list of references used to generate the validity summary tables.

Organizing the Summary Tables

Research Setting. As a first step in organizing the validity data, we decided that results should be reported according to the type of research setting, military versus non-military. We reasoned that this distinction might reveal differences in the type of predictors used, the type of criterion measures employed, and observed correlations in the two settings. This distinction was based on whether the subject population was military or civilian.

Validity data reported in the Military tables include results for predictors administered to Army, Marine Corps, Air Force, and Navy personnel. Data were obtained from a total of 27 technical reports and journal articles and represent a summary of approximately 2,900 coefficients.

Validities reported in Non-military tables have been summarized for predictors administered in private and public work settings and in high school or vocational-technical school settings. These data were obtained from 33 technical reports, journal articles, and test manuals and include more than 1,900 correlations between predictor and criterion measures.

Predictor Category. Within the Military and Non-military settings, validity data are organized by cognitive ability construct, using the nine broad cognitive ability constructs identified in Table 7 (Spatial, Perceptual Speed and Accuracy, Verbal, Reasoning, Number Facility, Memory, Perception, Fluency, and Mechanical Aptitude). Also included in the tables are three technical knowledge constructs. As noted previously, these are included to ensure that all information being assessed by the ASVAB is represented in the summary tables. The three technical knowledges are: (a) Electronics Information; (b) Auto, Shop, and Tool Knowledge; and (c) Science Knowledge.

Appendix C describes predictor measures included in each construct area. It is important to note that not all predictor measures included in the reported studies are described in that appendix, because complete descriptions of measures were not provided in all documents. Only those predictors for which authors provided complete test information are included in appendix C. These descriptions demonstrate the variety of measures that have been used to tap abilities in each of the predictor areas.

In this appendix, we have grouped validity predictor measures into the subcategories identified in Table 7 for the cognitive ability constructs. Although the validity data are not summarized by subcategory area, test descriptions are presented in this way to demonstrate how the measures may be sorted into those identified subcategories.

Criterion Category. Validity data were also categorized according to type of criterion measure. Researchers participating in the review process collaborated to identify and categorize criterion measures appearing in the literature. Table 28 provides a list and brief description of the criterion

Table 28

Criterion Constructs

Major Category	Criterion Construct	Definition or Explanation
Educational and School Achievement	Grades	Academic course grades or GPA
	Instructor Evaluations	Instructor ratings or rankings
Training Performance	Objective Measures	Paper-and-pencil exam scores, achievement test scores, or course grades based solely on paper-and-pencil exams
	Subjective Measures	Instructor ratings or rankings
	Combination Objective and Subjective Measures	Final course grades based on paper-and-pencil test scores and instructor evaluations. (Note: Unless it was specifically stated that training course grades were based on objective exams or subjective evaluations, they were categorized into this "combination" construct.)
	Go-No Go Training	Pass/fail, graduate/non-graduate, or successful/unsuccessful outcomes or number of washbacks
	Hands-On Measures	Work sample or job sample measures that are scored objectively or based on instructor evaluations
Job Proficiency	Ratings	Supervisor or peer ratings or rankings
	Job Knowledge Measures	Job knowledge or work sample tests
	Archival Measures	Units produced, salary rates or increases, or promotions

(Continued)

Table 28 (Continued)

Criterion Constructs

Major Category	Criterion Construct	Definition or Explanation
Job Involvement/ Withdrawal	Job Satisfaction	Job satisfaction or attitude survey ratings
	Job Withdrawal	Absenteeism, re-enlistment, or voluntary turnover
Adjustment[a]	Substance Abuse	Reported chemical abuse in a work setting
	Delinquency	Reported work-related problems such as Article 15 and AWOL
	Discharge Conditions	Unfavorable or dishonorable discharge from service

[a]Constructs in this category were geared toward situations that arise in the military. Although cognitive measures have been used to predict these work-related outcomes, it was expected that non-cognitive measures would be more effective for predicting scores on these constructs.

category areas: (a) Educational, (b) Training, (c) Job Proficiency, (d) Job Involvement, and (e) Adjustment. Within the cognitive area, we located only a few correlations between cognitive ability measures and Adjustment criterion measures, and found no correlations between cognitive ability measures and Job Involvement criterion measures. Thus, the bulk of the summarized validity coefficients involve Educational, Training, and Job Proficiency criterion measures.

Within each major criterion category, subcategories are listed and defined. Distinctions among these subcategories are clear-cut, with the possible exception of the three Training measures. To ensure consistency in classifying these criterion measures, we formulated the following guidelines: (a) Objective criteria include scores on periodic quizzes and final examinations; (b) subjective criteria include instructors' evaluations or ratings of students' performance in training; and (c) combination criteria include both examination scores, such as scores from quizzes or tests, and the instructor's evaluation of performance.

The Combination category also contains criterion measures that were described only as "final course grades." This decision was based on the assumption that final grades, unless otherwise specified, most likely include both objective and subjective components. In general, a large portion of training criteria involving course grades fall into the Combination category.

Job Type. The final factor used to organize and summarize the data is job type. This classification scheme was derived by examining two fairly well-known job classification systems: (a) the Dictionary of Occupational Titles (Department of Labor, 1977), and (b) Ghiselli's General Occupational Classification Scheme (Ghiselli, 1966). These grouping systems, along with information about Army MOS, were used to generate a job classification scheme that allowed categorization of all occupations appearing in the literature while still retaining important distinctions among broad job types. For example, separate categories were retained for mechanical maintenance and electronics job types rather than collapsing these two into a single category as in the DOT (i.e., structural occupations).

Included in the Military validity summary tables are seven broad job types: (1) Professional, Technical, and Managerial; (2) Clerical; (3) Protective Service; (4) Service; (5) Mechanical and Structural Maintenance; (6) Electronics; and (7) Miscellaneous. Job type categories for the Non-military data include all of the above and one additional category, Industrial Occupations, which did not appear to be represented by any military jobs in our review. A list of the job types is provided in Table 29, along with samples of specific Military and Non-military jobs included in each category.

In developing this job classification system, we initially included Sales jobs. It became clear, however, that very few jobs of this sort were included in studies that we reviewed. Hence, validity coefficients for Sales jobs are not presented in these tables.

Procedures for Summarizing Validity Coefficients

Sorting the Studies. After identifying the nine broad predictor categories, the five criterion areas, and the seven or eight job types, we began sorting the studies into Military versus Non-military groups and then proceeded to summarize the data in each research setting. Basically one staff member worked with the Military data, and another worked with the Non-military data. Although working alone to summarize the data for the two research settings, they frequently conferred to clarify the decision rules for classifying data by predictors, criterion measures, and job types.

For all studies the following information was obtained from Predictor and Article review forms:

1. Predictor Construct
2. Test Title
3. Criterion Construct
4. Validity Coefficient
5. Type of Validity Computed
6. Sample Size
7. Job Type
8. Article Form number
9. Predictor Form number

Table 29

Job Types and Sample Jobs

Job Type	Sample Military Jobs	Sample Non-Military Jobs
Professional, Technical, and Managerial	Air Force Officer Pilot Navigator Intelligence	Manager, Supervisor, Foreman Engineer Health Care Professional Pilot Draftsman
Clerical	Office Clerk Administrative Specialist Personnel Specialist Communications Specialist	Secretary Office Clerk Switchboard/Keyboard Operator Telegrapher
Protective Services	Military Police Combat Soldier Infantryman General Enlisted Personnel Undifferentiated Apprentices	Police Trainee Security Guard Corrections Officer
Service	Food Service Medical Specialist	Food Service Medical, Dental Assistant Truck Driver
Mechanical and Structural Maintenance	Aircraft Mechanic Vehicle Mechanic Munitions Mechanic	Machinist Mechanic Carpenter Plumber Welder Appliance Repairman
Electronics	Electronics and Radio Repairman Radar Repairman Sonar Technician Surveillance Specialist Radio Operator	Electronics Repairman Electrical Technology Trainee
Industrial	None	Machine Operator Processor, Assembler, Bench Worker Iron Worker Coal Miner General Maintenance Worker
Miscellaneous	Submarine Trainee	Power Plant Operator

Computation of Summary Information. For each Predictor by Criterion by Job Type cell, we computed the median coefficient across all studies. Also for each cell, we tallied the number of independent studies from which the validities were obtained, the number of validity coefficients used to compute the median value, the number of different predictor measures, and the range of sample sizes in these studies. These values are reported in each cell, along with the median value.

Within the Military studies, authors sometimes reported validity coefficients corrected for restriction in range. In some cases, both corrected and uncorrected validities were reported (Thomas & Thomas, 1965; Thomas, 1970). On rare occasions, authors reported only corrected validity coefficients (Massey & Creagor, 1956). Median values for corrected validity estimates are provided in the summary tables. Note that for the Military tables only, split cells provide uncorrected median validity estimates on the left and corrected validity estimates on the right. (Special notes are included on all Military summary tables to indicate how these data are organized.)

In summarizing the data for the Non-military tables, we attempted to include results from Ghiselli's summary (Ghiselli, 1966). The format he used to summarize data, however, was not easily amenable to the format we had developed. Therefore, the Non-military summary tables also contain split cells with values on the left representing median validity coefficients obtained from our literature review and values on the right representing median values obtained from Ghiselli's review. (Again, special notes indicate how to interpret the split cells in the Non-military summary tables.)

On the following pages, the median validity coefficients obtained for each Job Type within Non-military and Military occupations are reported. Following this discussion, we provide a condensed summary of these data that combines Non-military and Military validity estimates.

VALIDITY DATA SUMMARY

In each of the tables that follow, we report the median validity coefficient (mdn $\underline{r}$), number of independent studies (K), number of validity coefficients (L), number of different predictor measures (M), and sample size or range of sample sizes (N range). In the title of each table, the total number of validity estimates located for that job type and research setting is indicated. Only the uncorrected validity coefficients identified in the literature search are included in this count (i.e., corrected validities and median values obtained from Ghiselli's review are not included in this count.)

Professional, Technical, and Managerial

Non-military. Table 30 contains the validity data for Non-military Professional, Technical, and Managerial occupations (e.g., manager, supervisor, and foreman). For this job type, Number Facility (.45),

Table 30

Validity Summary for Non-Military Professional, Managerial and Technical Jobs (N = 280 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Combination (Ghiselli)	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Ratings (Ghiselli)	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Spatial												
mdn r	.24		.16	.36	.21	.28			.18[1]	.22		
K	16		2	1	1	abb			11	2d,4e		
L	16		5	1	1	3			12	6		
M	2		1	1	1				3			
N range	43-474		47-86	37	55				9-113			
Perceptual Speed and Accuracy												
mdn r	.16		.27	.45	.04	.21			.19	.32		
K	14		2	1	3	abb			10	a,4b		
L	14		5	1	10	3			12	5		
M	2		1	1	3				4			
N range	43-474		47-87	37	55-2000				9-113			
Verbal												
mdn r	.26		.11	.13	.37				.31			
K	16		2	2	1				12			
L	16		5	2	1				18			
M	2		1	1	1				4			
N range	43-474		47-96	37-48	55				9-113			
Reasoning												
mdn r	.30		.33	.54	.28				.38			
K	7		3	3	4				9			
L	30		3	3	6				19			
M	3		3	3	3				4			
N range	43-474		71-95	37-48	9-1828				9-99			
Number Facility												
mdn r	.45		.36	.48	.28	.33			.23	.23		
K	16		2	2	6	bbb			12	cdd		
L	16		5	2	8	3			13	3		
M	2		1	1	2				3			
N range	43-474		47-95	37-48	55-1828				9-113			

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Note 1: When a cell is divided by a broken line (only for training-combination and job-proficiency ratings criteria), the right side contains data from Ghiselli, 1966. The first number is the median validity coefficient across the studies reviewed by Ghiselli and the last number is the number of independent studies he reviewed. The middle set of numbers designates sample sizes for the independent studies, and the number of studies having each. They are coded as follows: a = <100, b = 100-499, c = 500-999, d = 1,000-4,999, e = 5,000-9,999, and f = 10,000+.

Note 2: A = written job knowledge tests, certification; B = oral job knowledge tests; C = simulated work samples; D = differences between high and low AFQT scorers, E = quantitative measure (i.e., units produced), F = salary, pay grade; G = promotions, highest level achieved, H = daily differences; and, I = injury index and lost time due to accidents.

Table 30 (Continued)

Validity Summary for Non-Military Professional, Managerial and Technical Jobs
(N = 280 Validity Coefficients)

CRITERION MEASURE

	Educational		Training						Job Proficiency			
PREDICTOR CONSTRUCT	Course Grades	Instructor Rankings	Objective	Subjective	Combination		Go-No Go	Hands-On	Ratings		Job Knowledge Tests	Archival Production
Memory												
mdn r												
K												
L												
M												
N range												
Perception												
mdn r	.17		.25		.06	.25			.12	.25		
K	16		1		1	2a,3b			9	abe		
L	16		1		1	5			11	3		
M	2		1		1				4			
N range	43-474		95		55				9-113			
Fluency												
mdn r	.21		.86						.42			
K	6		1						1			
L	6		1						1			
M	1		1						1			
N range	43-474		30						9			
Mechanical Aptitude												
mdn r	.36				.30				.09	.23		-.156
K	9				5				3	a,2e		1
L	9				5				3	3		2
M	2				1				2			1
N range	34-2878				85-1451				18-208			66
Auto/Shop/Tool												
mdn r												
K												
L												
M												
N range												
Electronics Knowledge												
mdn r												
K												
L												
M												
N range												
Science Knowledge												
mdn r												
K												
L												
M												
N range												

K - number of independent studies L - number of validity coefficients included M - number of different predictor measures included
N range - the sample size or range of sample size

Mechanical Aptitude (.36), Reasoning (.30), Verbal (.26), and Spatial abilities (.24) predict success in educational settings.

For training criteria, Number Facility, Reasoning, and Spatial abilities appear effective (median validity coefficients are equal or greater than .16 for three training criterion measures). Perceptual Speed and Accuracy (PS&A) measures correlate fairly well with two of the training criteria, objective and subjective (.27 and .45 respectively). Also note, however, that while Fluency correlates very highly with objective training criteria, the coefficient (.86) was obtained from a single study with a small sample (N = 30).

For job proficiency criteria, the highest correlations appear for Reasoning (.38), Verbal (.31), and Number Facility (.23). The correlation for Fluency is also high (.42), but is based on a very small sample size. Values from Ghiselli's review indicate that PS&A (.32), Perception (.25), Mechanical Aptitude (.23), and Number Facility (.23) are effective predictors of performance ratings.

Military. Data summarized in Table 31 indicate that markedly fewer validity coefficients were located for Professional, Managerial, and Technical occupations in the military (e.g., intelligence personnel) than for non-military occupations; validity data were located for training criteria only. For combination criteria, the best predictors are Number Facility (.62), Spatial abilities (.48), Reasoning (.47), PS&A (.41), and Perception (.35). Median validities, in general, are lower for go-no go training criteria; the best predictors are Reasoning, Perception, Verbal ability, PS&A, and Spatial abilities (median values are equal to or greater than .18). For hands-on criteria, Number Facility, Spatial abilities, Reasoning, PS&A, and Perception appear most effective (median values are greater than .30).

Clerical

Non-military. Table 32 summarizes 534 validity coefficients obtained for Non-military Clerical occupations (e.g., keyboard operator). In the area of educational criteria, Number Facility (.54), Verbal ability (.38), Perception (.35), PS&A (.34), Reasoning (.30), and Spatial abilities (.26) are the most effective predictors. According to the validity coefficients located for training criterion measures, all predictors appear effective; these include Spatial abilities, PS&A, Verbal ability, Number Facility, and Perception (median values are equal to or greater than .24). According to Ghiselli's review, Spatial abilities, PS&A, Number Facility, Memory, Perception, and Mechanical Aptitude are effective predictors of training for clerical personnel (median values are equal to or greater than .32).

For job proficiency rating measures, the best predictors of success in clerical occupations are Number Facility, Verbal ability, PS&A, and Reasoning (median values are equal to or greater than .16). Results from Ghiselli's review indicate that Number Facility, PS&A, Perception, and Mechanical Aptitude are the best predictors of job proficiency rating criteria (median values are equal to .21).

Table 31

<u>Validity Summary for Military Professional, Technical and Managerial Jobs (N = 101 Validity Coefficients)</u>

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Spatial										
mdn r					.48	.18	.40			
K					1	3	1			
L					10	7	10			
M					5	5	5			
N range					65-120	46-507	65-120			
Perceptual Speed and Accuracy										
mdn r					.41	.18	.33			
K					1	2	1			
L					2	6	2			
M					1	4	1			
N range					65-120	245-507	65-120			
Verbal										
mdn r						.19				
K						1				
L						1				
M						1				
N range						507				
Reasoning										
mdn r					.47	.20	.38			
K					1	2	1			
L					4	8	4			
M					2	4	2			
N range					65-120	245-507	65-120			
Number Facility										
mdn r					.62		.42			
K					1		1			
L					2		2			
M					1		1			
N range					65-120		65-120			

K = number of independent studies L = number o' validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample ize

(Continued)

Table 31 (Continued)

Validity Summary for Military Professional, Technical and Managerial Jobs (N = 101 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Memory										
mdn r						.11				
K						2				
L						2				
M						1				
N range						245-507				
Perception										
mdn r					.35	.20	.31			
K					1	3	1			
L					14	3	14			
M					7	2	7			
N range					65-120	245-507	65-120			
Fluency										
mdn r										
K										
L										
M										
N range										
Mechanical Aptitude										
mdn r						.10				
K						3				
L						4				
M						2				
N range						46-507				
Auto/Shop/Tool										
mdn r						.08				
K						2				
L						4				
M						3				
N range						245-507				
Electronics Knowledge										
mdn r						.10				
K						2				
L						2				
M						1				
N range						245-507				
Science Knowledge										
mdn r										
K										
L										
M										
N range										

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

Table 32

Validity Summary for Non-Military Clerical Jobs (N = 534 Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Spatial										
mdn r	.26			.36	¦ .36			.07 ¦ .15	.01[C]	.14[E]
K	2			1	¦ a,6c,			9 ¦ 14a,14b	5	1
L	3			1	¦ 7d,4e,			10 ¦ 5c,9d,	5	1
M	2			1	¦ 5f			31[1] ¦ e	1	1
N range	51-312			59	¦ 23			203 ¦ 43	50-130	50
Perceptual Speed and Accuracy										
mdn r	.34			.29	.33 ¦ .35			.17 ¦ .21	.26[A,C]	.11[E,F,H]
K	2			3	3 ¦ 7a,12b			92 ¦ 14a,19b	9	10
L	3			3	3 ¦ 3c,3e			124 ¦ 8c,13d	33	12
M	2			2	1 ¦ 25			6 ¦ e	8	1
N range	51-312			59-173	40-173 ¦			20-103 ¦ 55	50-1091	23-177
Verbal										
mdn r	.38		.24	.55				.18	.27[A,C]	.15[E,F,H]
K	2		1	1				70	9	10
L	3		1	1				112	13	12
M	2		1	1				8	3	1
N range	51-312		40	59				20-390	50-191	23-177
Reasoning										
mdn r	.30							.16		
K	1							44		
L	2							45		
M	1							4		
N range	247-312							11-400		
Number Facility										
mdn r	.54			.34	.28 ¦ .50			.20 ¦ .21	.30[C]	.23[E,F,H]
K	1			3	3 ¦ 2a,c,			87 ¦ 4a,7b	5	10
L	1			3	3 ¦ 4d,4e			121 ¦ 2c,5b	13	12
M	1			2	1 ¦ 11			4 ¦ 2e,f	3	3
N range	51			59-173	40-173 ¦			20-390 ¦ 21	50-1091	23-177

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Note 1: When a cell is divided by a broken line (only for training-combination and job-proficiency ratings criteria), the right side contains data from Ghiselli, 1966. The first number is the median validity coefficient across the studies reviewed by Ghiselli and the last number is the number of independent studies he reviewed. The middle set of numbers designates sample sizes for the independent studies, and the number of studies having each. They are coded as follows: a = <100, b = 100-499, c = 500-999, d = 1,000-4,999, e = 5,000-9,999, and f = 10,000+.

Note 2: A = written job knowledge tests, certification; B = oral job knowledge tests; C = simulated work samples; D = differences between high and low AFQT scorers, E = quantitative measure (i.e., units produced), F = salary, pay grade; G = promotions, highest level achieved, H = daily differences; and, I = injury index and lost time due to accidents.

Table 32 (Continued)

Validity Summary for Non-Military Clerical Jobs (N = 534 Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Memory										
mdn r					.32			.08 / .14		
K					2a,4d			11 / 2a,7b		
L					6			11 / 2c,2d		
M								1 / e		
N range								99-350 / 14		
Perception										
mdn r	.35			.36	.34			.10 / .21	.27[C]	.17[E]
K	1			1	2a,7b			11 / 11a,7b	5	1
L	1			1	c,3e			13 / c,8d	5	1
M	1			1	13			3 / e,f	1	1
N range	51			59				31-203 / 29	50-130	50
Fluency										
mdn r								.10		
K								7		
L								7		
M								2		
N range								35-303		
Mechanical Aptitude										
mdn r	.16				.32			.06 / .21		
K	1				3c,d			1 / 3c,d		
L	2				2e,3f			1 / 2e,3f		
M	1				9			1 / 9		
N range	247-312							185		
Auto/Shop/Tool										
mdn r	.05									
K	1									
L	4									
M	2									
N range	247-312									
Electronics Knowledge										
mdn r	.11									
K	1									
L	2									
M	1									
N range	247-312									
Science Knowledge										
mdn r										
K										
L										
M										
N range										

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

Job knowledge tests are predicted best by measures of Number Facility, Verbal ability, PS&A, and Perception (median values are greater than .25). In predicting archival production scores, the best measures are Number Facility and Perception (median values are equal to or greater than .17).

Military. Table 33 contains median validity coefficients for Clerical occupations (e.g., administrative specialist). For uncorrected values, Verbal ability, Mechanical Aptitude, Reasoning, PS&A, Electronics Knowledge, Spatial abilities, Memory, Perception, and Auto/Shop/Tool are effective predictors of success measured by objective training criteria (median values range from .23 to .56). Most of these validity coefficients represent values obtained from a single study. Values for corrected coefficients suggest that Science Knowledge and Number Facility are also useful predictors of training criteria. For combination training criterion measures, the highest uncorrected validities appear for Electronics Knowledge (.36), Number Facility (.11), Verbal ability (.30), Memory (.29), and Reasoning (.27). Validity estimates for go-no go criterion measures are much lower with median values ranging from .05 to .11.

For job proficiency criteria, correlations between predictors and ratings range from -.14 (Perception) to .10 (Memory). Values for job knowledge tests range from .01 (Fluency) to .36 (Verbal ability); the highest uncorrected values appear for Verbal ability, Auto/Shop/Tool, Electronics Knowledge, Number Facility, and Memory (median *r* values are equal to or greater than .19). Corrected values indicate that Reasoning, Number Facility, Science Knowledge, Electronics Knowledge, Mechanical Aptitude, PS&A, Auto/Shop/Tool Knowledge, and Spatial abilities are effective predictors of job knowledge tests (median values are equal to or greater than .34).

Protective Services

Non-military. We located only a few validity estimates for this job type in a non-military setting (e.g., corrections officer). Note that most of the estimates appearing in Table 34 were obtained from Ghiselli's review. According to his summary, Mechanical Aptitude, Number Facility, Spatial abilities, PS&A, Perception, and Memory are effective predictors of training criteria (median *r* for all predictors is equal to or greater than .28).

According to the data we located, PS&A, Verbal ability, Number Facility, and Perception are effective predictors of job proficiency ratings (median values are equal to or greater than .19). Results from Ghiselli's review indicate that Mechanical Aptitude (.29) and Fluency (.26) are also effective predictors of job ratings.

Military. Table 35 presents validity estimates for Protective Service occupations in the military (e.g., infantryman). For training criteria (uncorrected validities), it appears that all measures used are fairly successful in predicting combination training scores (median validity coefficients range from .26 to .47); no Combination data were located for Spatial abilities, Perception, or Fluency. The best predictors of hands-on training scores are Verbal ability (.18), Memory (.17), and Reasoning (.15).

Table 33

Validity Summary for Military Clerical Jobs (N = 620 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective (uncorrected)	Training: Objective (corrected)	Training: Subjective	Training: Combination (uncorrected)	Training: Combination (corrected)	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests (uncorrected)	Job Proficiency: Job Knowledge Tests (corrected)	Job Proficiency: Archival Production
Spatial													
mdn r			.34D	.27						.05	.13A	.34A	
K			1	1						1	1	1	
L			3	4						8	16	3	
M			3	1						2	2	1	
N range			201	77-311						193-234	57-102	87-421	
Perceptual Speed and Accuracy													
mdn r			.37D	.26	.13	.16	.23	.10		.06	.02A	.38A	
K			1	1	1	6	2	1		2	1	1	
L			2	4	12	65	29	12		13	12	6	
M			2	1	6	14	1	6		4	3	2	
N range			201	77-311	225-232	61-1551		225-232		54-234	57-102	87-421	
Verbal													
mdn r			.56D	.50	.16	.30	.45	.11		.01	.36A		
K			1	1	1	5	2	1		1	1		
L			1	4	2	45	29	2		8	8		
M			1	1	1	4	1	1		2	2		
N range			201	77-311	225-232	38-1551	61-1551	225-232		193-234	57-102		
Reasoning													
mdn r			.40D	.49		.27	.47			.08	.13A	.60A	
K			1	1		3	2			2	1	1	
L			4	4		75	29			14	24	3	
M			4	1		5	1			5	3	1	
N range			201	77-311		65-1551				54-234	57-102	87-421	
Number Facility													
mdn r				.54		.31				.08	.22A	.57A	
K				1		2				1	1	1	
L				8		9				4	8	3	
M				2		1				1	1	1	
N range				77-311		95-362				193-234	57-102	87-421	

K = number of independent studies   L = number of validity coefficients included   M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Note 1: Split cells in these tables indicate corrected and uncorrected validity coefficients are provided. Values on the left are uncorrected; values on the right are corrected for restriction of range.

Note 2: A = written job knowledge tests, certification; B = oral job knowledge tests; C = simulated work samples; D = differences between high and low AFQT scorers, E = quantitative measure (i.e., units produced), F = salary, pay grade; G = promotions, highest level achieved, H = daily differences; and, I = injury index and lost time due to accidents.

Table 33 (Continued)

Validity Summary for Military Clerical Jobs (N = 620 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Memory										
mdn r			.26[D]		.29			.10	.19[A]	
K			1		1			2	1	
L			1		1			6	8	
M			1		1			3	1	
N range			206		95			54-234	57-102	
Perception										
mdn r			.25[A]	.11	.14	.05		-.14		
K			1	1	1	1		1		
L			2	4	4	4		3		
M			2	2	2	2		3		
N range			201	225-232	225-232	225-232		54		
Fluency										
mdn r								.05	.01[A]	
K								1	1	
L								4	8	
M								1	1	
N range								193-234	57-102	
Mechanical Aptitude										
mdn r			.44[D] .33		.16 .20			.03	.03[A] .45[A]	
K			1 1		4 2			1	1 1	
L			1 4		38 27			4	8 3	
M			1 1		2 1			1	1 1	
N range			201 77-311		61-1551			193-234	57-102 87-421	
Auto/Shop/Tool										
mdn r			.23[D] .14		.19 .37			.00	.26[A] .35[A]	
K			1 1		1 1			1	1 1	
L			2 8		19 19			4	8 6	
M			2 2		1 1			1	1 2	
N range			201 77-311		63-1551			193-234	57-102 87-421	
Electronics Knowledge										
mdn r			.35[D] .35		.36 .50			.00	.23[A] .49[A]	
K			1 1		1 1			1	1 1	
L			1 4		19 19			4	8 3	
M			1 1		1 1			1	1 1	
N range			201 77-311		63-1551			193-234	57-102 87-421	
Science Knowledge										
mdn r			.47						.53[A]	
K			1						1	
L			4						3	
M			1						1	
N range			77-311						87-421	

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

Table 34

Validity Summary for Non-Military Protective Services Jobs (N = 8 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination		Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings		Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Spatial												
mdn r						.32			-.15	.16		
K						5b,c			1	2a,4b		
L						6			1	c		
M									1	7		
N range									51			
Perceptual Speed and Accuracy												
mdn r						.30			.38	.18		
K						b			1	4b		
L						1			1	4		
M									1			
N range									51			
Verbal												
mdn r									.28			
K									1			
L									1			
M									1			
N range									51			
Reasoning												
mdn r												
K												
L												
M												
N range												
Number Facility												
mdn r						.38			.25	.01		
K						bcc			1	a,3b		
L						3			1	4		
M									1			
N range									51			

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Note 1: When a cell is divided by a broken line (only for training-combination and job-proficiency ratings criteria), the right side contains data from Ghiselli, 1966. The first number is the median validity coefficient across the studies reviewed by Ghiselli and the last number is the number of independent studies he reviewed. The middle set of numbers designates sample sizes for the independent studies, and the number of studies having each. They are coded as follows: a = <100, b = 100-499, c = 500-999, d = 1,000-4,999, e = 5,000-9,999, and f = 10,000+.

Table 34 (Continued)

Validity Summary for Non-Military Protective Services Jobs (N = 8 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination		Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings		Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Memory												
mdn r						.28						
K						b						
L						1						
M												
N range												
Perception												
mdn r						.30			.19	.20		
K						b			1	bb		
L						1			1	2		
M									1			
N range									51			
Fluency												
mdn r										.26		
K										a,2b		
L										3		
M												
N range												
Mechanical Aptitude												
mdn r			.20	.10	.28	.45		.27		.29		
K			1	1	1	bbb		1		a,3b		
L			1	1	1	3		1		4		
M			1	1	1			1				
N range			129	129	129			129				
Auto/Shop/Tool												
mdn r												
K												
L												
M												
N range												
Electronics Knowledge												
mdn r												
K												
L												
M												
N range												
Science Knowledge												
mdn r												
K												
L												
M												
N range												

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

Table 35

Validity Summary for Military Protective Services Jobs (N = 323 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Combination (corrected)	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests	Job Proficiency: Job Knowledge Tests (corrected)	Job Proficiency: Archival Production
Spatial												
mdn r								.04	.13	-.08	.37[A]	
K								2	2	2	1	
L								15	3	12	8	
M								3	2	3	1	
N range								42-142	299-718	29-211	93-2733	
Perceptual Speed and Accuracy												
mdn r				.31	.27	.22		.13	.11	.01	.24[A]	
K				1	3	2		3	4	2	1	
L				1	23	11		23	7	15	8	
M				1	6	1		6	4	3	1	
N range				146	74-595	114-595		42-146	70-694	29-211	93-2733	
Verbal												
mdn r				.15	.31	.48		.18	.13			.09[F]
K				1	2	2		2	3			1
L				4	13	13		11	4			1
M				4	1	1		5	3			1
N range				146-148	114-595			140-148	70-624			7,923
Reasoning												
mdn r				.18	.42	.45		.15	.17		.51[A]	.06[F]
K				1	3	2		3	4		1	1
L				1	19	13		11	7		8	1
M				1	4	1		3	5		1	1
N range				146	74-595	114-595		42-146	70-773		93-2733	7,923
Number Facility												
mdn r				.34	.34			.13			.45[A]	
K				1	1			2			1	
L				2	2			10			16	
M				2	1			2			2	
N range				146	74-223			97-146			93-2733	

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Note 1: Split cells in these tables indicate corrected and uncorrected validity coefficients are provided. Values on the left are uncorrected; values on the right are corrected for restriction of range.

Note 2: A = written job knowledge tests, certification; B = oral job knowledge tests; C = simulated work samples; D = differences between high and low AFQT scorers, E = quantitative measure (i.e., units produced), F = salary, pay grade; G = promotions, highest level achieved, H = daily differences; and, I = injury index and lost time due to accidents.

Table 35 (Continued)

Validity Summary for Military Protective Services Jobs (N = 323 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Memory										
mdn r					.31		.17	.13	-.04C	
K					1		2	3	2	
L					2		6	4	5	
M					1		1	4	1	
N range					74-233		42-142	111-356	29-211	
Perception										
mdn r							.08	.12	.08C	
K							3	4	2	
L							12	8	10	
M							3	8	2	
N range							42-142	111-755	29-211	
Fluency										
mdn r										
K										
L										
M										
N range										
Mechanical Aptitude										
mdn r				.20	.26 / .46		.11	.13	.06B / .51A	
K				1	3 / 2		2	3	1 / 1	
L				1	15 / 13		5	4	1 / 8	
M				1	2 / 1		1	2	1 / 1	
N range				146	74-595 / 114-595		97-146	70-624	29-38 / 93-2733	
Auto/Shop/Tool										
mdn r				.30	.31 / .53		.11	.12	.43A	
K				1	1 / 1		2	3	1	
L				1	8 / 8		8	8	16	
M				1	1 / 1		4	4	2	
N range				146	114-595		97-146	70-624	93-2733	
Electronics Knowledge										
mdn r				.23	.47 / .66		.04	.09	.50A	
K				1	1 / 1		2	1	1	
L				1	8 / 8		5	4	8	
M				1	1 / 1		1	2	1	
N range				146	114-595		97-146	497-624	93-2733	
Science Knowledge										
mdn r				.23			.08		.50A	
K				1			2		1	
L				1			5		8	
M				1			1		1	
N range				146			97-146		93-2733	

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

Number Facility and PS&A appear somewhat useful in predicting these criteria (median _r_ = .13).

For job proficiency ratings, median validities range from .09 (Electronics Knowledge) to .17 (Reasoning), with most of the correlations at .12 or .13. Uncorrected correlations computed between predictors and job knowledge test scores are all low (range -.08 to .08). According to the corrected values, however, Mechanical Aptitude, Reasoning, Number Facility, Spatial abilities, PS&A, and measures in the three technical knowledge areas are effective predictors of job knowledge test scores (values range from .24 to .51). Results on this table also suggest that Verbal ability and Reasoning have been used to predict pay grade; these correlations are, however, very low.

Service

Non-military. Table 36 contains median validity estimates for Service occupations (e.g., medical or dental assistant). For educational criteria, Perception, PS&A, Spatial abilities, Reasoning, Verbal ability, and Electronics Knowledge appear to be the best predictors (median values are equal to or greater than .29).

For objective and subjective training criterion measures, Number Facility, Spatial ability, PS&A, Perception appear to be the best predictors (median values are equal to or greater than .22). Spatial abilities and Perception are the best predictors of the combination training criteria. Results from Ghiselli's review indicate that Number Facility (.54), Spatial ability (.42), and Mechanical Aptitude (.36) are the best predictors of training criteria.

For job proficiency criterion measures, the best predictors are Spatial abilities (.27), Number Facility (.25), Perception (.24), and Mechanical Aptitude (.21). According the Ghiselli's summary, Memory should be added to this list (median _r_ = .29).

Military. Table 37 contains the median validity estimates for military Service occupations (e.g., food service specialist). According to the uncorrected median estimates for training criteria, Verbal Ability (.47), Electronics Knowledge (.46), Number Facility (.32), and Reasoning (.30) are the best predictors. Given the corrected values, Auto/Shop/Tool, Mechanical Aptitude, and PS&A could be added to this list (median corrected values are equal to or greater than .30).

Median validities are low for job proficiency rating criterion measures, ranging from .00 to .17. The highest values appear for PS&A and Memory. For job knowledge tests, median uncorrected values range from .04 to .49, with the highest values appearing for Electronics Knowledge, Verbal ability, Auto/Shop/Tool, Number Facility, and Spatial abilities (median values are greater than .30). Median values for corrected validities indicate that Science Knowledge, Mechanical Aptitude, and Reasoning are also effective predictors of job knowledge test scores.

Table 36

Validity Summary for Non-Military Service Jobs (N = 130 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Combination (Ghiselli)	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Ratings (Ghiselli)	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Spatial												
mdn r	.35		.35	.31	.25	.42			.27	.15		
K	2		1	4	1	4c,d			13	abcd		
L	4		1	4	1	5			13	4		
M	2		1	1	1				1			
N range	31-163		85	46-85	160				31-121			
Perceptual Speed and Accuracy												
mdn r	.36		.22	.35	.10	.19			.17	.18		
K	2		1	4	1	c			12	a,5b		
L	4		1	4	1	1			12	6		
M	2		1	1	1				1			
N range	31-163		81	46-85	60				31-121			
Verbal												
mdn r	.29		.24	.16	.12				.18			
K	2		1	4	1				13			
L	4		1	4	1				15			
M	2		1	1	1				3			
N range	31-163		81	46-85	160				31-121			
Reasoning												
mdn r	.32								-.02			
K	1								1			
L	3								1			
M	1								1			
N range	31-163								60			
Number Facility												
mdn r	.22		.41	.34	.12	.54			.25	.04		
K	1		1	4	1	ccd			13	bbb		
L	1		1	4	1	3			14	3		
M	1		1	1	1				2			
N range	65		81	46-85	160				31-121			

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Note 1: When a cell is divided by a broken line (only for training-combination and job-proficiency ratings criteria), the right side contains data from Ghiselli, 1966. The first number is the median validity coefficient across the studies reviewed by Ghiselli and the last number is the number of independent studies he reviewed. The middle set of numbers designates sample sizes for the independent studies, and the number of studies having each. They are coded as follows: a = <100, b = 100-499, c = 500-999, d = 1,000-4,999, e = 5,000-9,999, and f = 10,000+.

Table 36 (Continued)

Validity Summary for Non-Military Service Jobs (N = 130 Validity Coefficients)

CRITERION MEASURE

	Educational		Training						Job Proficiency			
PREDICTOR CONSTRUCT	Course Grades	Instructor Rankings	Objective	Subjective	Combination		Go-No Go	Hands-On	Ratings		Job Knowledge Tests	Archival Production
Memory												
mdn r						.10			.06	.29		
K						d			1	b		
L						1			1	1		
M									1			
N range									60			
Perception												
mdn r	.38		.22	.28	.20	.08			.24	.04		
K	1		1	4	1	c			13	bbb		
L	1		1	4	1	1			13	3		
M	1		1	1	1				1			
N range	65		81	46-86	160				31-121			
Fluency												
mdn r												
K												
L												
M												
N range												
Mechanical Aptitude												
mdn r	.24					.36			.21	.21		
K	1					d			2	d		
L	3					1			2	1		
M	1								2			
N range	31-163								60-155			
Auto/Shop/Tool												
mdn r	.19											
K	1											
L	6											
M	2											
N range	31-163											
Electronics Knowledge												
mdn r	.29											
K	1											
L	3											
M	1											
N range	31-163											
Science Knowledge												
mdn r												
K												
L												
M												
N range												

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

Table 37

Validity Summary for Military Service Jobs (N = 237 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination		Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests		Job Proficiency: Archival Production
Spatial												
mdn r				.21	.23				.07	.31[A]	.45[A]	
K				1	[illegible]				1	1	1	
L				4	4				4	8	4	
M				4	4				2	2	1	
N range				137	95				247-261	50-116	93-405	
Perceptual Speed and Accuracy												
mdn r					.20	.30			.17	.08[A]	.25[A]	
K					4	2			1	1	1	
L					33	8			6	18	4	
M					6	1			3	3	1	
N range					111-5325	293-5323			247-261	50-116	93-405	
Verbal												
mdn r					.47	.58			.06	.40[A]	.48[A]	
K					2	2			1	1	1	
L					8	8			4	6	4	
M					1	1			2	2	1	
N range					293-5325				247-261	50-116	93-405	
Reasoning												
mdn r					.30	.54			.05	.16[A]	.48[A]	
K					4	2			1	1	1	
L					51	8			6	12	4	
M					5	1			3	3	1	
N range					111-5325	293-5325			247-261	50-116	93-405	
Number Facility												
mdn r					.32				.09	.34[A]	.42[A]	
K					2				1	1	1	
L					11				2	4	8	
M					1				1	1	2	
N range					111-996				247-261	50-116	93-405	

K = number of independent studies   L = number of validity coefficients included   M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Note 1: Split cells in these tables indicate corrected and uncorrected validity coefficients are provided. Values on the left are uncorrected; values on the right are corrected for restriction of range.

Note 2: A = written job knowledge tests, certification; B = oral job knowledge tests; C = simulated work samples; D = differences between high and low AFQT scorers, E = quantitative measure (i.e., units produced), F = salary, pay grade; G = promotions, highest level achieved, H = daily differences; and, I = injury index and lost time due to accidents.

Table 37 (Continued)

Validity Summary for Military Service Jobs (N = 237 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Combination	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Memory												
mdn r					.21				.14	.27A		
K					1				1	1		
L					1				2	4		
M					1				1	1		
N range					389				247-261	50-116		
Perception												
mdn r				.21	.06							
K				1	1							
L				1	1							
M				1	1							
N range				137	95							
Fluency												
mdn r									.01	.04A		
K									1	1		
L									2	4		
M									1	1		
N range									247-261	50-116		
Mechanical Aptitude												
mdn r					.21	.31			.00	.21A	.51	
K					4	2			1	1	1	
L					19	8			2	4	4	
M					2	1			1	1	1	
N range					111-5325	293-5325			247-261	50-116	93-405	
Auto/Shop/Tool												
mdn r					.21	.41			.05	.36A	.50	
K					1	1			1	1	1	
L					4	4			2	4	8	
M					1	1			1	1	2	
N range					293-5325				247-261	50-116	93-405	
Electronics Knowledge												
mdn r					.46	.60			.08	.49A	.55A	
K					1	1			1	1	1	
L					4	4			2	4	4	
M					1	1			1	1	1	
N range					293-5325				247-261	50-116	93-405	
Science Knowledge												
mdn r											.52A	
K											1	
L											4	
M											1	
N range											93-405	

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

Mechanical and Structural Maintenance

Non-military. Median validity estimates for non-military Mechanical and Structural Maintenance occupations are presented in Table 38 (e.g., carpenter, plumber, welder). For course grade criterion measures in educational settings, Number Facility, Reasoning, and Mechanical Aptitude are the best predictors (median values are greater than .25). Note that instructor rankings yield negative correlations with scores on many of the predictor measures. The best predictors are Perception (.36) and PS&A (.23), but these data were obtained from only one or two studies.

For training criteria, the most effective predictors are Mechanical Aptitude, Perception, Verbal ability, Number Facility, PS&A, and Spatial abilities (median values are equal to or greater than .23). Results from Ghiselli's review suggest that these same measures effectively predict training criteria, with the exception of Verbal ability for which no data are available.

For job proficiency rating criteria, Number Facility, Mechanical Aptitude, PS&A, and Spatial abilities are the most effective predictors (median values are equal to or greater than .17). Median values from Ghiselli's summary would add Memory and Perception to this list of predictors of job ratings. In addition, Verbal ability (.46) and Electronics Knowledge (.38) predict archival production scores and Mechanical Aptitude (.23) and PS&A (.20) predict adjustment scores for this occupation.

Military. Table 39 presents median validity estimates for military Mechanical and Structural Maintenance occupations (e.g., light wheel vehicle mechanic). For training course grades (objective, subjective, and combination criterion measures), the best predictors are Electronics Knowledge, Mechanical Aptitude, Auto/Shop/Tool, and Verbal ability (median values are greater than .25). Corrected validity estimates for these criterion measures indicate that Reasoning and Number Facility are also effective predictors. For go-no go criterion measures, median values range from .02 to .22, with Verbal ability the best predictor. Median values range from .05 to .15 for hands-on criterion measures; Mechanical Aptitude is the best predictor of this criterion measure.

Median correlations computed between predictor scores and job proficiency ratings range from -.09 (Reasoning) to .12 (Mechanical Aptitude). For job knowledge tests, median values are higher, ranging from .04 to .45. For this criterion measure, uncorrected validities indicate that Auto/Shop/Tool, Verbal ability, Mechanical Aptitude, and Spatial abilities are the best predictors. Focusing on the corrected validity estimates, Science Knowledge, Number Facility, and Reasoning should be added to the list of effective predictors.

Electronics

Non-military. Table 40 presents median validity estimates for non-military Electronics occupations (e.g., electronics repairman). For educational criteria, the most effective predictors are Number Facility,

Table 38

<u>Validity Summary for Non-Military Mechanical and Structural Maintenance Jobs</u>
<u>(N = 491 Validity Coefficients)</u>

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination		Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings		Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production	Adjustment: Delinquency
Spatial													
mdn r	.22	-.22		.34	.23	.23			.17	.20		-.01 I	.04
K	6	2		4	25	6a,7b			54	4a,12b		1	1
L	22	4		4	28	4c,2d			56	2c,7d		3	3
M	3	4		1	3	f			7	e		3	3
N range	20-322	48-74		40-63	42-75	20			47-480	26		48-74	48-74
Perceptual Speed and Accuracy													
mdn r	.17	.23		.27	.27	.30			.18	.17		.01 I	.20
K	6	2		3	4	4a,6b			33	18a,5b		1	1
L	22	2		3	4	6c,7d			43	3c,d		1	1
M	3	2		1	1	4e,8f			5	27		1	1
N range	20-322	49-58		40-63	42-75	35			49-350			49	49
Verbal													
mdn r	.22	-.06		.26	.34				.11			.46 E,I	.02
K	9	2		5	4				31			2	1
L	27	2		6	4				33			2	1
M	3	2		2	1				6			2	1
N range	20-322	58-116		25-63	42-75				25-350			53-116	116
Reasoning													
mdn r	.40	-.61		.05					.14			-.35 I	.01
K	2	1		1					23			1	1
L	10	1		1					22			1	1
M	4	1		1					6			1	1
N range	20-322	67		58					25-350			67	67
Number Facility													
mdn r	.41	-.04		.48	.28	.36			.19	.20		.19 E,I	-.02
K	8	2		6	5	3c,2d			26	6b,2c		2	1
L	10	2		7	5	4e,4f			31	2d		2	1
M	2	2		3	2	13			7	10		2	1
N range	22-66	58-116		25-63	42-75				25-350			53-116	116

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Note 1: When a cell is divided by a broken line (only for training-combination and job-proficiency ratings criteria), the right side contains data from Ghiselli, 1966. The first number is the median validity coefficient across the studies reviewed by Ghiselli and the last number is the number of independent studies he reviewed. The middle set of numbers designates sample sizes for the independent studies, and the number of studies having each. They are coded as follows: a = <100, b = 100-499, c = 500-999, d = 1,000-4,999, e = 5,000-9,999, and f = 10,000+.

Note 2: A = written job knowledge tests, certification; B = oral job knowledge tests; C = simulated work samples; D = differences between high and low AFQT scorers, E = quantitative measure (i.e., units produced), F = salary, pay grade; G = promotions, highest level achieved, H = daily differences; and, I = injury index and lost time due to accidents.

Table 38 (Continued)

Validity Summary for Non-Military Mechanical and Structural Maintenance Jobs
(N = 491 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination		Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings		Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production	Adjustment: Delinquency
Memory													
mdn r									.07	.17			
K									3	a,b,c			
L									3	3			
M									1				
N range									74-233				
Perception													
mdn r	.12	.36		.32	.40	.30			.13	.17			
K	5	1		3	4	3a,5b			30	5a,3b			
L	5	1		3	4	5c,2d			35	4c,2d			
M	2	1		1	1	3f			4	2f			
N range	36-66	58		40-63	42-75	18			49-480	16			
Fluency													
mdn r	.16	-.13							.05			-.10	.01
K	1	1							5			1	1
L	1	1							5			1	1
M	1	1							2			1	1
N range	36	64							60-350			64	64
Mechanical Aptitude													
mdn r	.26	-.15			.44	.39			.19	.19		.09[E,J]	.23
K	1	1			2	2b,3c			16	5b,2c		2	1
L	17	1			2	2d,e			21	2d		2	1
M	1	1			1	5f			6	9		2	1
N range	20-322	132			75-155	13			47-480			134-430	132
Auto/Shop/Tool													
mdn r	.21												
K	1												
L	14												
M	2												
N range	20-322												
Electronics Knowledge													
mdn r	.21								.08			.38[G]	
K	1								4			1	
L	7								7			1	
M	1								1			1	
N range	20-322								47-480			430	
Science Knowledge													
mdn r													
K													
L													
M													
N range													

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

Table 39

Validity Summary for Military Mechanical and Structural Maintenance Jobs
(N = 594 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective		Training: Subjective	Training: Combination		Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests		Job Proficiency: Archival Production
Spatial													
mdn r				.28						-.05	.33[A]	.29[A]	
K				1						1	1	1	
L				2						4	8	3	
M				1						2	2	1	
N range			124-128							236-264	72-95	131-659	
Perceptual Speed and Accuracy													
mdn r				.21	.08	.15	.22	.07	.05	-.04	.04[AB]	.21[AB]	
K				1	1	5	2	1	1	1	2	2	
L				2	6	92	63	6	2	6	6	5	
M				1	6	12	1	6	1	3	5	3	
N range			124-128		230	33-6934		230	49-50	236-264	72-186	131-659	
Verbal													
mdn r				.39	.27	.34	.53	.22	.09	.00	.43[AC]	.30[AB]	
K				1	1	4	2	1	1	1	2	2	
L				2	1	80	64	1	2	4	7	4	
M				1	1	4	1	1	1	2	3	2	
N range			124-128		230	33-6934		230	49-50	236-264	72-186	131-659	
Reasoning													
mdn r				.39	.15	.23	.50	.09	.12	-.09	.14[AB]	.30[AB]	
K				1	1	6	2	1	1	1	2	2	
L				2	1	104	63	1	2	6	13	4	
M				1	1	7	1	1	1	3	4	2	
N range			124-128		230	33-6934		230	49-50	236-264	72-186	131-659	
Number Facility													
mdn r				.34		.23				-.06	.20[A]	.48[A]	
K				1		2				1	2	2	
L				4		10				2	5	7	
M				2		1				1	2	2	
N range			124-128			81-1147				236-264	72-186	131-659	

K = number of independent studies   L = number of validity coefficients included   M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Note 1: Split cells in these tables indicate corrected and uncorrected validity coefficients are provided. Values on the left are uncorrected; values on the right are corrected for restriction of range.

Note 2: A = written job knowledge tests, certification; B = oral job knowledge tests; C = simulated work samples; D = differences between high and low AFQT scorers, E = quantitative measure (i.e., units produced), F = salary, pay grade; G = promotions, highest level achieved, H = daily differences; and, I = injury index and lost time due to accidents.

Table 39 (Continued)

Validity Summary for Military Mechanical and Structural Maintenance Jobs (N = 594 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Memory										
mdn r					.11			-.07	$.11^{AB}$	
K					1			1	2	
L					1			2	5	
M					1			1	2	
N range					315			236-264	72-186	
Perception										
mdn r				.07	.10	.02				
K				1	1	1				
L				2	2	2				
M				2	1	2				
N range				230	230	230				
Fluency										
mdn r								-.01	.15	
K								1	1	
L								2	4	
M								1	1	
N range								236-264	72-95	
Mechanical Aptitude										
mdn r			.41		.25 ¦ .46		.15	.12	$.35^{ABC}$ ¦ $.51^{A}$	
K			1		4 ¦ 2		1	1	4 ¦ 1	
L			2		73 ¦ 63		2	2	10 ¦ 3	
M			1		2 ¦ 1		1	1	4 ¦ 1	
N range			124-128		33-6934		49-50	236-264	72-461 ¦ 131-659	
Auto/Shop/Tool										
mdn r			.41		.32 ¦ .50		.10	.09	$.45^{A}$ ¦ $.53^{AB}$	
K			1		1 ¦ 1		1	1	1 ¦ 2	
L			4		39 ¦ 39		2	2	4 ¦ 7	
M			2		1 ¦ 1		1	1	1 ¦ 3	
N range			124-128		90-6934		49-50	236-264	72-95 ¦ 131-659	
Electronics Knowledge										
mdn r			.27		.39 ¦ .56			-.03	$.26^{AB}$ ¦ $.25^{AB}$	
K			1		2 ¦ 2			1	2 ¦ 2	
L			2		42 ¦ 42			2	5 ¦ 5	
M			1		1 ¦ 1			1	2 ¦ 2	
N range			124-128		90-6934			236-264	186-326 ¦ 131-659	
Science Knowledge										
mdn r			.37						$.49^{A}$	
K			1						1	
L			2						3	
M			1						1	
N range			124-128						131-659	

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

Table 40

Validity Summary for Non-Military Electronics Jobs (N = 130 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Spatial										
mdn r	.15		.61	.55			.30	.24		
K	4		1	3			1	9		
L	5		1	3			1	9		
M	2		1	2			1	2		
N range	63-392		90	10-50			90	50-216		
Perceptual Speed and Accuracy										
mdn r	.19		.34	.47			.10	.21		
K	4		1	3			1	8		
L	5		1	3			1	12		
M	2		1	2			1	4		
N range	63-392		90	10-50			90	50-216		
Verbal										
mdn r	.27		.55	.18			.31	.22		
K	8		1	3			1	9		
L	10		1	3			1	10		
M	3		1	2			1	3		
N range	22-392		90	10-50			90	50-216		
Reasoning										
mdn r	.23		.65	.50			.43	.01		
K	1		1	2			1	1		
L	2		3	2			3	1		
M	1		3	2			3	1		
N range	292-392		90	10-18			90	101		
Number Facility										
mdn r	.32	.60	.38				.39	.19		
K	7	1	3				1	9		
L	8	1	3				1	9		
M	2	1	2				1	2		
N range	22-127	90	10-50				90	51-216		

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Note 1: A = written job knowledge tests, certification; B = oral job knowledge tests; C = simulated work samples; D = differences between high and low AFQT scorers, E = quantitative measure (i.e., units produced), F = salary, pay grade; G = promotions, highest level achieved, H = daily differences; and, I = injury index and lost time due to accidents.

Table 40 (Continued)

Validity Summary for Non-Military Electronics Jobs (N = 130 Validity Coefficients)

CRITERION MEASURE

	Educational		Training					Job Proficiency		
PREDICTOR CONSTRUCT	Course Grades	Instructor Rankings	Objective	Subjective	Combination	Go-No Go	Hands-On	Ratings	Job Knowledge Tests	Archival Production
Memory										
mdn r										
K										
L										
M										
N range										
Perception										
mdn r	.03		.47	.13				.25		
K	3		1	1				9		
L	3		2	1				10		
M	1		1	1				3		
N range	63-127		90	90				50-216		
Fluency										
mdn r			.27				.13			
K			1				1			
L			1				1			
M			1				1			
N range			90				90			
Mechanical Aptitude										
mdn r	.23								.32[A]	
K	1								3	
L	2								3	
M	1								2	
N range	292-312								101-216	
Auto/Shop/Tool										
mdn r	.23									
K	1									
L	4									
M	2									
N range	292-392									
Electronics Knowledge										
mdn r	.29								.27[A]	
K	1								2	
L	2								2	
M	1								1	
N range	292-392								101-216	
Science Knowledge										
mdn r										
K										
L										
M										
N range										

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

Electronics Knowledge, and Verbal ability (median values are greater than .25).

For training criterion measures (objective and subjective), the most effective predictors are Spatial abilities, Reasoning, Number Facility, and PS&A (median values are .34 or greater). Based on data from single studies, Verbal ability and Perception are also effective predictors of objective training measures. For hands-on measures, the best predictors are Reasoning, Number Facility, Verbal ability, and Spatial abilities (values are equal to or greater than .30). Note that these validity estimates were obtained from single studies.

The most effective predictors of job proficiency ratings are Perception, Spatial abilities, Verbal ability, PS&A, and Number Facility (median values are greater than .19). Correlations with job knowledge tests were found for only two predictor constructs, Mechanical Aptitude (.32) and Electronics Knowledge (.27).

Military. Median validity estimates are summarized for military Electronics personnel in Table 41 (e.g., radar repairman). For training criteria, uncorrected median validities indicate that Electronics Knowledge, Verbal ability, Reasoning, and Auto/Shop/Tool knowledge are the best predictors (median values are greater than .25). Data available for corrected validity estimates suggest that, in addition to the predictors listed, Mechanical Aptitude, Science Knowledge, and PS&A correlate highly with training scores.

Median correlations between predictors and job proficiency ratings range from .01 to .13, with the highest values appearing for Number Facility (.13) and Auto/Shop/Tool knowledge (.11). Median uncorrected validity estimates for job knowledge tests range from -.01 to .37, with the highest values appearing for Electronics Knowledge (.37) and Verbal ability (.34). According to the median values computed for corrected validities, Electronics Knowledge, Mechanical Aptitude, Science Knowledge, Reasoning, Number Facility, Verbal ability, and PS&A are effective predictors of job knowledge test scores (median values are greater than .25).

Industrial

Non-military. Table 42 presents median validity estimates for non-military Industrial occupations (e.g., machine operator). For educational criteria, Spatial abilities (.60), Perception (.41), PS&A (.36), and Number Facility (.35) are the most effective predictors. Note that most of these estimates are obtained from single studies.

For training criteria, Number Facility, Perception, Verbal ability, PS&A, and Spatial abilities all appear effective (median values are equal to or greater than .24). According to results from Ghiselli's review, Mechanical Aptitude would also be added to this list of predictors.

For job proficiency ratings, Mechanical Aptitude, Perception, Spatial abilities, Number Facility, Memory, and PS&A yield median correlations of

Table 41

Validity Summary for Military Electronics Jobs (N = 130 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Spatial										
mdn r			.27		.38			.07	.18[A]	
K			1		1			2	2	
L			4		1			24	44	
M			1		1			2	2	
N range			78-257		108			97-416	27-174	
Perceptual Speed and Accuracy										
mdn r			.28		.09 / .26			.08	-.01[A] / .26[A]	
K			1		3 / 2			3	3 / 1	
L			4		74 / 60			36	56 / 2	
M			1		3 / 1			3	3 / 1	
N range			78-257		55-1175			27-416	27-174 / 147-202	
Verbal										
mdn r			.38		.39 / .65			.05	.34[A] / .28[B]	
K			1		4 / 2			2	2 / 2	
L			4		65 / 60			22	44 / 3	
M			1		3 / 1			2	3 / 1	
N range			78-257		53-1175			97-416	27-174 / 147-202	
Reasoning										
mdn r			.44		.30 / .54	.48		.07	.14[A] / .47[A]	
K			1		4 / 2	1		3	2 / 1	
L			4		91 / 60	2		40	66 / 2	
M			1		6 / 1	2		5	3 / 1	
N range			78-257		53-1175	144-173		27-416	27-174 / 147-202	
Number Facility										
mdn r			.47	.12				.13	.17[A] / .44	
K			1	1				2	2 / 1	
L			8	7				12	22 / 4	
M			2	1				1	1 / 2	
N range			78-257	99-138				97-416	27-174 / 147-202	

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Note 1: Split cells in these tables indicate corrected and uncorrected validity coefficients are provided. Values on the left are uncorrected; values on the right are corrected for restriction of range.

Note 2: A = written job knowledge tests, certification; B = oral job knowledge tests; C = simulated work samples; D = differences between high and low AFQT scorers, E = quantitative measure (i.e., units produced), F = salary, pay grade; G = promotions, highest level achieved, H = daily differences; and, I = injury index and lost time due to accidents.

Table 41 (Continued)

Validity Summary for Military Electronics Jobs (N = 130 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Memory										
mdn r								.02	.10[A]	
K								3	2	
L								16	22	
M								3	1	
N range								27-416	27-174	
Perception										
mdn r								.04		
K								1		
L								6		
M								3		
N range								27-37		
Fluency										
mdn r								.01	.06	
K								2	2	
L								12	22	
M								1	1	
N range								97-416	27-174	
Mechanical Aptitude										
mdn r			.43		.17 .37			.06	.22 .53	
K			1		3 2			2	2 1	
L			4		67 60			12	22 2	
M			1		2 1			1	1 1	
N range			78-257		53-11735			97-416	27-174 147-202	
Auto/Shop/Tool										
mdn r			.35		.27 .43			.11	.22	
K			1		1 1			2	2	
L			8		35 35			12	22	
M			2		1 1			1	1	
N range			78-257		53-11735			97-416	97-174	
Electronics Knowledge										
mdn r			.44		.40 .61			.02	.37 .54[A]	
K			1		2 2			2	2 1	
L			4		57 57			12	22 2	
M			1		1 1			1	1 1	
N range			78-257		53-11735			97-416	27-174 147-202	
Science Knowledge										
mdn r			.43						.51[A]	
K			1						1	
L			4						2	
M			1						1	
N range			78-257						147-202	

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

Table 42

Validity Summary for Non-Military Industrial Jobs (N = 308 Validity Coefficients)

CRITERION MEASURE

PREDICTOR CONSTRUCT	Educational: Course Grades	Educational: Instructor Rankings	Training: Objective	Training: Subjective	Training: Combination	Training: Go-No Go	Training: Hands-On	Job Proficiency: Ratings	Job Proficiency: Ratings (Ghiselli)	Job Proficiency: Job Knowledge Tests	Job Proficiency: Archival Production
Spatial											
mdn r	.60			.24	.36			.29	.23		
K	1			5	2a,2b			41	13a,17b		
L	1			5	5c,18d			41	9c,12d		
M	1			2	2e			3	5e,f		
N range	59			20-164	29			31-327	57		
Perceptual Speed and Accuracy											
mdn r	.36			.31	.24			.24	.20		
K	1			5	5a,8b			42	28a		
L	1			5	6c,3d			44	15b,c		
M	1			2	23			5	9d,e		
N range	59			20-164				14-327	53		
Verbal					*				*		
mdn r	.24			.35	.35			.18	.22		.32[E]
K	2			5	64-70			50	64-70		6
L	2			5	1			57	2		9
M	2			2				6			2
N range	59-153			20-164				31-327			48-178
Reasoning											
mdn r								.22			
K								7			
L								8			
M								2			
N range								31-234			
Number Facility											
mdn r	.35			.38	.39			.27	.22		.25[E]
K	2			5	2a,3b			51	6a,11b		5
L	2			5	2c,2d			56	3d		9
M	2			2	e			4	20		2
N range	59-153			20-164	10			31-327			48-178

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Note 1: When a cell is divided by a broken line (only for training-combination and job-proficiency ratings criteria), the right side contains data from Ghiselli, 1966. The first number is the median validity coefficient across the studies reviewed by Ghiselli and the last number is the number of independent studies he reviewed. The middle set of numbers designates sample sizes for the independent studies, and the number of studies having each. They are coded as follows: a = <100, b = 100-499, c = 500-999, d = 1,000-4,999, e = 5,000-9,999, and f = 10,000+.

Note 2: A = written job knowledge tests, certification; B = oral job knowledge tests; C = simulated work samples; D = differences between high and low AFQT scorers, E = quantitative measure (i.e., units produced), F = salary, pay grade; G = promotions, highest level achieved, H = daily differences; and, I = injury index and lost time due to accidents.

* = Cell includes the validity coefficient, number of independent studies, and number of coefficients reported.

Table 42 (Continued)

Validity Summary for Non-Military Industrial Jobs (N = 308 Validity Coefficients)

CRITERION MEASURE

	Educational		Training					Job Proficiency			
PREDICTOR CONSTRUCT	Course Grades	Instructor Rankings	Objective	Subjective	Combination	Go-No Go	Hands-On	Ratings		Job Knowledge Tests	Archival Production
Memory											
mdn r					.12			.24	.15		
K					bbb			3	8a,7b		
L					3			3	3d		
M								2	18		
N range								14-234			
Perception											
mdn r	.41			.36	.27			.30	.20		-.46E
K	1			3	5a,4b			41	15a,12b		1
L	1			3	8c,d			42	4c,5d		1
M	1			1	2e			5	2e,f		1
N range	59			50-164	20			31-327	39		90
Fluency											
mdn r								.19			
K								1			
L								1			
M								1			
N range								234			
Mechanical Aptitude											
mdn r	.18				.40			.35	.40		
K	2				c,4d			5	3a,7b		
L	2				5			5	c		
M	1							2	11		
N range	511-634							14-662			
Auto/Shop/Tool											
mdn r											
K											
L											
M											
N range											
Electronics Knowledge											
mdn r											
K											
L											
M											
N range											
Science Knowledge											
mdn r											
K											
L											
M											
N range											

K - number of independent studies L - number of validity coefficients included M - number of different predictor measures included
N range - the sample size or range of sample size

.24 or greater. For the most part, Ghiselli's summary indicates validities lower than those we located. For archival production scores, Verbal ability and Number Facility correlate positively, while Perception correlates negatively.

Miscellaneous

Non-military. Table 43 presents median validity values for non-military miscellaneous occupations (e.g., power plant operator). For educational criteria, PS&A, Number Facility, and Spatial abilities are the best predictors (median values range from .24 to .30).

Correlations between predictor scores and job proficiency ratings range from .02 to .21. The most effective predictors are Verbal ability (.21), Reasoning (.18), and Spatial abilities (.17).

Military. Median validity coefficients for military miscellaneous occupations are presented in Table 44 (e.g., Submarine trainee). Note that validities were located for only 4 of the 12 predictors and all validities were obtained from a single study. These data indicate that Number Facility, Electronics Knowledge, Reasoning, and Science Knowledge are effective predictors of objective training criteria (median values are equal to or greater than .25). Median validity estimates for subjective and combination criterion measures are low or negative. For hands-on criteria, Number Facility is the best predictor (.24)

Summary of Military and Non-military Validity Tables

The purpose of the validity summary is to identify cognitive ability predictors that might be used to supplement the current military selection and classification battery, the ASVAB. In organizing the summary tables, we also planned to examine differences between data reported in military versus non-military settings. These differences are discussed below.

First, from the summary tables it is clear that measures of technical knowledge have been widely used in all military branches. In fact, these types of measures had been used well before the ASVAB was implemented DOD-wide in 1976. It is also apparent from the military summary tables that such measures have been useful in predicting training and job performance outcomes for a variety of MOS. It is clear from the non-military tables that measures of technical knowledge have been used much less often in private business and school settings. The one exception to this finding is the predictor construct, Mechanical Aptitude. Recall that we elected to include this measure in the cognitive construct taxonomy because it appears useful for a wide variety of occupations.

Second, correlations between predictors and job proficiency ratings differ, on the average, for the two research settings. In military settings, the median values across all predictor constructs and across all job types are very low or near zero; the median value across all studies, predictor constructs, and job types is .06. Median correlations between predictors and job proficiency ratings reported in non-military studies are higher than those observed for military settings; the median value across

Table 43

Validity Summary for Non-Military Miscellaneous Jobs (N = 69 Validity Coefficients)

CRITERION MEASURE

	Educational		Training					Job Proficiency		
PREDICTOR CONSTRUCT	Course Grades	Instructor Rankings	Objective	Subjective	Combination	Go-No Go	Hands-On	Ratings	Job Knowledge Tests	Archival Production
Spatial										
mdn r	.24							.17		
K	1							1		
L	1							39		
M	1							2		
N range	54							185-3158		
Perceptual Speed and Accuracy										
mdn r	.30							.10		
K	1							1		
L	1							2		
M	1							2		
N range	54							74		
Verbal										
mdn r	.01							.21		
K	1							2		
L	1							19		
M	2							2		
N range	54							74-2051		
Reasoning										
mdn r								.18		
K								1		
L								1		
M								1		
N range								74		
Number Facility										
mdn r	.29							.09		
K	1							1		
L	1							1		
M	1							1		
N range	54							74		

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Table 43 (Continued)

Validity Summary for Non-Military Miscellaneous Jobs (N = 69 Validity Coefficients)

CRITERION MEASURE

	Educational		Training						Job Proficiency	
PREDICTOR CONSTRUCT	Course Grades	Instructor Rankings	Objective	Subjective	Combination	Go-No Go	Hands-On	Ratings	Job Knowledge Tests	Archival Production
Memory										
mdn r										
K										
L										
M										
N range										
Perception										
mdn r	.16									
K	1									
L	1									
M	1									
N range	54									
Fluency										
mdn r										
K										
L										
M										
N range										
Mechanical Aptitude										
mdn r								.02		
K								1		
L								1		
M								1		
N range								74		
Auto/Shop/Tool										
mdn r										
K										
L										
M										
N range										
Electronics Knowledge										
mdn r								.05		
K								1		
L								1		
M								1		
N range								74		
Science Knowledge										
mdn r										
K										
L										
M										
N range										

K = number of independent studies   L = number of validity coefficients included   M = number of different predictor measures included
N range = the sample size or range of sample size

Table 44

Validity Summary for Military Miscellaneous Jobs (N = 24 Validity Coefficients)

CRITERION MEASURE

	Educational		Training					Job Proficiency		
PREDICTOR CONSTRUCT	Course Grades	Instructor Rankings	Objective	Subjective	Combination	Go-No Go	Hands-On	Ratings	Job Knowledge Tests	Archival Production
Spatial										
mdn r										
K										
L										
M										
N range										
Perceptual Speed and Accuracy										
mdn r										
K										
L										
M										
N range										
Verbal										
mdn r										
K										
L										
M										
N range										
Reasoning										
mdn r			.32	.10	.04		.10			
K			1	1	1		1			
L			2	2	2		2			
M			2	2	2		2			
N range			76	76	76		76			
Number Facility										
mdn r			.50	.00	-.20		.24			
K			1	1	1		1			
L			1	1	1		1			
M			1	1	1		1			
N range			76	76	76		76			

K = number of independent studies L = number of validity coefficients included M = number of different predictor measures included
N range = the sample size or range of sample size

(Continued)

Table 44 (Continued)

Validity Summary for Military Miscellaneous Jobs (N = 24 Validity Coefficients)

CRITERION MEASURE

	Educational		Training					Job Proficiency		
PREDICTOR CONSTRUCT	Course Grades	Instructor Rankings	Objective	Subjective	Combination	Go-No Go	Hands-On	Ratings	Job Knowledge Tests	Archival Production
Memory										
mdn r										
K										
L										
M										
N range										
Perception										
mdn r										
K										
L										
M										
N range										
Fluency										
mdn r										
K										
L										
M										
N range										
Mechanical Aptitude										
mdn r										
K										
L										
M										
N range										
Auto/Shop/Tool										
mdn r										
K										
L										
M										
N range										
Electronics Knowledge										
mdn r			.41	-.05	-.23		-.12			
K			1	1	1		1			
L			2	2	2		2			
M			2	2	2		2			
N range			76	76	76		76			
Science Knowledge										
mdn r			.25	.01	.03		.07			
K			1	1	1		1			
L			1	1	1		1			
M			1	1	1		1			
N range			76	76	76		76			

K - number of independent studies L - number of validity coefficients included M - number of different predictor measures included
N range - the sample size or range of sample size

all studies, predictor constructs, and job types is .18. According to results from Ghiselli's review, the median value for job ratings is .20 across all predictors and job types.

Validity data for job rating criterion measures, then, indicate that measures used to capture job performance via ratings differ for military and non-military settings. Reasons for these differences are unclear; it may be due to differences in job structure. For example, soldiers are required to demonstrate job skills in both garrison and field settings. Supervisors may differ for the two settings, thereby preventing them (raters) from observing and evaluating a soldier in all job areas. The variations may also be related to the broader definition of job performance in the military. That is, job performance may encompass not only technical requirements of a particular job but also general soldiering skills, military bearing and appearance, and adjustment factors. Based on the limited amount of data reported in the summary tables, we would expect the correlation between cognitive ability measures and adjustment measures to be low or near zero.

It would be useful to investigate the source of these differences to understand why cognitive ability constructs appear more predictive of job performance ratings in non-military than in military settings. The design of the current Project A allows comparison of validities computed using different types of rating measures. For example, while rating scales are being constructed to assess specific MOS performance requirements, separate scales are being developed to assess general soldier performance requirements, such as military bearing, leadership abilities, and adjustment. Results from analyses using these distinct types of performance rating scales may yield higher correlations between cognitive ability predictors and ratings of technical job performance than correlations between cognitive predictors and general soldier performance ratings. If, indeed, military job proficiency rating scales described in the literature have confounded job performance and "general soldier" requirements, we would expect to find that validities computed using MOS-specific job performance rating scales are nearly as high as those observed in the non-military literature.

A final distinction between military and non-military studies involves the use of archival data to predict cognitive ability test scores. This particular criterion construct includes such things as units produced, salary or pay grade, promotions or highest level achieved, injury index, and lost time due to accidents. In the military literature, we located only one study in which correlations were computed using this type of criterion measure. Many more studies employing this criterion measure were located in the Non-military literature. Overall, these data suggest that for Clerical, Mechanical and Structural Maintenance, and Industrial occupations, cognitive ability measures may predict archival criterion scores.

SECTION SUMMARY AND CONCLUSIONS

Because it is difficult to succinctly summarize the validity data presented in the foregoing tables, we have generated yet another table (Table 45) that presents median values for military and non-military data combined. This table differs from Tables 30 to 44 in several ways. First, median values are reported only for uncorrected validity coefficients that we located in our review of the literature (i.e., corrected median validity coefficients and values summarized by Ghiselli are not included in this table).

Second, median values are reported only for the four broad criterion measures. Thus, for a particular predictor construct, we computed the median value for Educational criteria, which include both course grades and instructor rankings. For Training criteria, we computed the median value for a particular predictor across objective, subjective, combination, go-no go, and hands-on training measures. For Job Proficiency measures, ratings, job knowledge tests, and archival data were combined to estimate the median validity for a single predictor construct. For Adjustment measures, we presented data for the small number of studies available.

Also note in this table that in each row (predictor construct) median values are reported for the eight job types and for All Job Types combined. Each column (job type) contains median values for the four criterion categories and a final Overall median value. The only additional information included is the number of validity coefficients used to compute the median value; this number is presented in parentheses. Because the focus of the current project is on predicting training and job performance outcomes, results for those two criterion categories are emphasized in the discussion that follows.

According to the data in Table 45, Spatial ability measures are effective predictors of Training outcomes for Electronics, Professional/Technical/Managerial, Clerical, Service, Mechanical and Structural Maintenance, and Industrial occupations (median values range from .24 to .49 with the Overall median at .26). The Overall value across all job types for Training is .26. For Job Proficiency criteria, Spatial ability measures are effective for Industrial, Service, Professional/Technical/Managerial, Mechanical and Structural Maintenance, and Miscellaneous, occupations (median values range from .17 to .25, with the Overall median value at .16).

Measures of Perceptual Speed and Accuracy appear to be effective predictors of Training criteria in Industrial, Service, Protective Service, Professional/Technical/Managerial, and Clerical occupations (median values range from .16 to .31 with the Overall median value at .16). For Job Proficiency criteria, measures of PS&A appear most effective for Industrial, Professional/Technical/Managerial, and Clerical occupations (median values range from .16 to .24, with the Overall value equal to .13).

Verbal ability is an effective predictor of Training outcomes in nearly all occupational groups. Values range from .16 (Professional/Technical/Managerial) to .35 (Electronics and Industrial), with the median Overall value equal to .31. Median validity estimates computed across all Job Proficiency criteria are somewhat lower than those for Training, but are still relatively high for all job types. Values range from .15 (Protective Services) to .31 (Professional/Technical/Managerial), with a median Overall value of .21.

Table 45

Median Validities for Cognitive Ability and Knowledge Constructs Summarized by Job Type and Criterion Category

JOB TYPE

COGNITIVE ABILITIES & CRITERION CATEGORIES	Professional, Technical & Managerial		Clerical		Protective Services		Service		Mechanical & Structural Maintenance		Electronics		Industrial		Miscellaneous		All Job Types	
Spatial																		
Education	.24	(16)	.26	(3)	--	--	.35	(4)	.15	(26)	.15	(5)	.60	(1)	.24	(1)	.21	(56)
Training	.34	(34)	.35	(4)	.04	(15)	.26	(14)	.24	(32)	.49	(6)	.24	(5)	--	--	.26	(110)
Job Proficiency	.18	(12)	.08	(40)	-.04	(16)	.25	(25)	.17	(71)	.15	(77)	.29	(41)	.17	(39)	.16	(321)
Adjustment	--	--	--		--	--	--	--	.04	(3)	--	--	--	--	--	--	.04	(3)
Overall	.28	(62)	.11	(47)	.00	(31)	.26	(43)	.18	(132)	.18	(88)	.29	(47)	.17	(40)	.19	(490)
Perceptual Speed and Accuracy																		
Education	.16	(14)	.34	(3)	--	--	.36	(4)	.18	(24)	.19	(5)	.36	(1)	.30	(1)	.20	(52)
Training	.18	(26)	.16	(97)	.20	(47)	.21	(39)	.15	(113)	.11	(79)	.31	(5)	--	--	.16	(406)
Job Proficiency	.19	(12)	.16	(194)	.06	(23)	.13	(36)	.14	(56)	.05	(104)	.24	(44)	.10	(2)	.13	(470)
Adjustment	--	--	--	--	--	--	--	--	.20	(1)	--	--	--	--	--	--	.20	(1)
Overall	.18	(52)	.16	(294)	.15	(70)	.18	(79)	.15	(194)	.08	(188)	.25	(50)	.17	(3)	.15	(929)
Verbal																		
Education	.26	(16)	.38	(3)	--	--	.29	(4)	.20	(29)	.27	(10)	.24	(2)	.01	(1)	.24	(65)
Training	.16	(9)	.30	(52)	.24	(28)	.34	(14)	.33	(94)	.35	(70)	.35	(5)	--	--	.31	(272)
Job Proficiency	.31	(18)	.19	(153)	.15	(6)	.21	(25)	.18	(46)	.24	(76)	.20	(66)	.21	(19)	.21	(409)
Adjustment	--	--	--	--	--	--	--	--	.02	(1)	--	--	--	--	--	--	.02	(1)
Overall	.26	(43)	.22	(208)	.22	(34)	.26	(43)	.27	(170)	.29	(156)	.21	(73)	.20	(20)	.25	(747)
Reasoning																		
Education	.30	(30)	.30	(2)	--	--	.32	(3)	.31	(11)	.23	(2)	--	--	--	--	.30	(48)
Training	.33	(28)	.28	(79)	.32	(31)	.30	(51)	.22	(109)	.32	(101)	--	--	.14	(8)	.28	(407)
Job Proficiency	.38	(19)	.14	(83)	.16	(8)	.12	(19)	.09	(36)	.11	(107)	.22	(18)	.18	(1)	.14	(281)
Adjustment	--	--	--	--	--	--	--	--	.01	(1)	--	--	-	--	--	--	.01	(1)
Overall	.35	(77)	.21	(164)	.29	(39)	.25	(73)	.20	(157)	.21	(210)	.22	(18)	.14	(9)	.23	(737)
Number Facility																		
Education	.45	(16)	.54	(1)	--	--	.22	(1)	.34	(12)	.35	(9)	.35	(2)	.29	(1)	.38	(42)
Training	.33	(19)	.31	(15)	.19	(12)	.32	(17)	.32	(22)	.22	(11)	.38	(5)	.14	(4)	.29	(105)
Job Proficiency	.23	(13)	.21	(158)	.25	(1)	.25	(20)	.18	(40)	.16	(43)	.27	(65)	.09	(1)	.21	(341)
Adjustment	--	--	--	--	--	--	--	--	-.02	(1)	--	--	--	--	--	--	-.02	(1)
Overall	.34	(48)	.22	(174)	.19	(13)	.24	(38)	.24	(75)	.20	(63)	.28	(72)	.16	(6)	.34	(489)
Memory																		
Education	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Training	.11	(2)	.28	(2)	.21	(8)	.21	(1)	.11	(1)	--	--	--	--	--	--	.20	(14)
Job Proficiency	--	--	.12	(25)	.04	(9)	.20	(7)	.06	(10)	.07	(38)	.24	(3)	--	--	.10	(92)
Adjustment	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Overall	.11	(2)	.13	(27)	.12	(17)	.20	(8)	.07	(11)	.07	(38)	.24	(3)	--	--	.16	(106)

(Continued)

Table 45 (Continued)

Median Validities for Cognitive Ability and Knowledge Constructs Summarized by Job Type and Criterion Category

COGNITIVE ABILITIES & CRITERION CATEGORIES	JOB TYPE: Professional, Technical & Managerial		Clerical		Protective Services		Service		Mechanical & Structural Maintenance		Electronics		Industrial		Miscellaneous		All Job Types	
Perception																		
Education	.17	(16)	.35	(1)	--	--	.38	(1)	.16	(6)	.03	(3)	.41	(1)	.16	(1)	.17	(29)
Training	.31	(33)	.14	(15)	.08	(12)	.23	(8)	.23	(13)	.36	(13)	.36	(3)	--	--	.25	(97)
Job Proficiency	.12	(11)	.11	(22)	.10	(19)	.24	(13)	.13	(35)	.17	(16)	.28	(13)	--	--	.18	(159)
Adjustment	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Overall	.26	(60)	.13	(38)	.09	(31)	.24	(22)	.16	(54)	.18	(32)	.29	(47)	.16	(1)	.20	(285)
Fluency																		
Education	.21	(6)	--	--	--	--	--	--	.02	(2)	--	--	--	--	--	--	.16	(8)
Training	.86	(1)	--	--	--	--	--	--	--	--	.20	(2)	--	--	--	--	.42	(3)
Job Proficiency	.42	(1)	.05	(19)	--	--	.03	(6)	.06	(12)	.04	(34)	.19	(1)	--	--	.05	(73)
Adjustment	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Overall	.31	(8)	.05	(19)	--	--	.03	(6)	.05	(14)	.05	(36)	.19	(1)	--	--	.07	(84)
Mechanical Aptitude																		
Education	.36	(9)	.16	(2)	--	--	.24	(3)	.24	(18)	.23	(2)	.18	(2)	--	--	.26	(36)
Training	.21	(9)	.17	(39)	.22	(25)	.21	(19)	.25	(77)	.17	(67)	--	--	--	--	.21	(236)
Job Proficiency	-.01	(5)	.03	(13)	.12	(5)	.16	(8)	.23	(35)	.18	(37)	.35	(5)	.02	(1)	.17	(109)
Adjustment	--	--	--	--	--	--	--	--	.23	(1)	--	--	--	--	--	--	.23	(1)
Overall	.22	(23)	.13	(54)	.20	(30)	.24	(30)	.24	(131)	.17	(106)	.30	(7)	.02	(1)	.20	(382)
Auto/Shop/Tool Knowledge																		
Education	--	--	.05	(4)	--	--	.19	(6)	.21	(14)	.23	(4)	--	--	--	--	.19	(28)
Training	.08	(4)	.19	(21)	.22	(17)	.21	(4)	.31	(45)	.27	(35)	--	--	--	--	.27	(126)
Job Proficiency	--	--	.17	(12)	.12	(8)	.26	(6)	.33	(6)	.18	(34)	--	--	--	--	.19	(66)
Adjustment	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Overall	.08	(4)	.17	(37)	.18	(25)	.22	(16)	.29	(65)	.23	(73)	--	--	--	--	.23	(220)
Electronics Knowledge																		
Education	--	--	.11	(2)	--	--	.29	(3)	.21	(7)	.29	(2)	--	--	--	--	.22	(14)
Training	--	--	.36	(20)	.30	(14)	.46	(4)	.39	(42)	.40	(57)	--	--	.00	(8)	.38	(145)
Job Proficiency	.10	(2)	.15	(12)	.09	(4)	.35	(6)	.15	(15)	.25	(36)	--	--	.05	(1)	.21	(76)
Adjustment	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Overall	.10	(2)	.28	(34)	.25	(18)	.37	(13)	.31	(64)	.26	(95)	--	--	.01	(9)	.27	(235)
Science Knowledge																		
Education	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Training	--	--	--	--	.11	(6)	--	--	--	--	--	--	--	--	.09	(4)	.10	(10)
Job Proficiency	--	--	--	--	--	--	--	--	--	--	.51	(2)	--	--	--	--	.51	(2)
Adjustment	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Overall	--	--	--	--	.11	(6)	--	--	--	--	.51	((2)	--	--	.09	(4)	.17	(12)

Measures assessing Reasoning abilities are effective predictors of Training outcomes for nearly all occupations, ranging from .14 (Miscellaneous) to .33 (Professional/Technical/Managerial), with an Overall value of .28. For Job Proficiency criteria, measures of Reasoning abilities are most effective for Professional/Technical/Managerial, Industrial, and Protective Services occupations (median values for these occupations range from .16 to .38, with an Overall median across all job types equal to .14).

Number Facility measures appear effective for predicting success in training for nearly all occupational groups; median values range from .14 (Miscellaneous) to .38 (Industrial), with an Overall median of .29. Note that across all job types, we located many more validities for Job Proficiency criterion measures (n=341) than for Training measures (n=105). Median validities for Job Proficiency measures are somewhat lower than those for Training measures. For this criterion, values range from .09 (Miscellaneous) to .27 (Industrial), with an Overall value of .21.

There were fewer validity coefficients located for measures of Memory relative to other cognitive ability constructs. According to these limited data, measures of this construct are effective for predicting Training criteria in Clerical (.28), Protective Services (.21), and Service (.21) occupations with an Overall median value of .20. Median values for Job Proficiency criteria indicate that Memory is most effective for Industrial (.24) and Service (.20) occupations, with an Overall median of .10.

Measures of perceptual abilities (Perception) are effective predictors of Training criteria for Electronics, Industrial, Professional/Technical/Managerial, Service, and Mechanical and Structural Maintenance occupations (median values for these job types range from .23 to .36 with the Overall median across all job types at .25). For Job Proficiency criterion measures, Perception tests are most effective for Industrial (.28), Service (.24), and Electronics (.17) occupations with an Overall median across all job types of .18.

According to the data reported in Table 45, it is fairly uncommon for Fluency measures to appear in either military or non-military validity studies. In fact, we located only three coefficients for Training criteria. For Professional/Technical/Managerial occupations, a single study was located; the resulting value (.86) is based on a small sample. For Electronics occupations, two validity coefficients suggest that Fluency may be an effective predictor of Training outcomes. For Job Proficiency criteria, most median values are low, with the Overall across all job types at .05.

Measures of Mechanical Aptitude are effective predictors of Training success for Mechanical and Structural Maintenance, Protective Services, Service, Professional/Technical/Managerial, Clerical, and Electronics occupations (median values range from .17 to .25 with an Overall median value of .21). For Job Proficiency criteria, measures of this construct are effective for Industrial (.35), Mechanical and Structural Maintenance (.23), Electronics (.18), and Service (.16) occupations, with the Overall median across all job types at .17.

Auto, Shop, and Tool knowledge measures are effective predictors of training outcomes for Mechanical and Structural Maintenance (.31), Electronics (.27), Protective Services (.22), Service (.21), and Clerical (.19) occupations, with an Overall median across all job types of .27. For Job Proficiency criteria, these knowledge measures are most effective in predicting success in Mechanical and Structural Maintenance (.33), Service (.26), Electronics (.18), and Clerical (.17) occupations, with the Overall median across all jobs at .19.

Electronics Knowledge measures, although generally reserved for military selection and classification purposes, effectively predict Training success for Service, Electronics, Mechanical and Structural Maintenance, Clerical, and Protective Services occupations (median values for these job types range from .30 to .46 with the Overall median across all job types at .38). For Job Proficiency criterion measures, Electronics Knowledge tests are most effective for Service (.35) and Electronics (.25) occupations; the Overall median is .21.

The final measure included in Table 45 is Science Knowledge. Because very few validities were located for this measure, these results are difficult to interpret.

In general, these summary data indicate that nearly all of the cognitive ability constructs included in the taxonomy are effective for predicting training or job performance success in one or more of the broad job categories. In the final part of this report we examine the implications of these data for constructing predictor measures to supplement the current military selection and classification battery. Before we begin that discussion, some observations about the data summarized in Table 45 are warranted.

First, note that for most job types, median validity estimates are higher, on the average, for Training criteria than for Job Proficiency criteria. Across the 12 cognitive ability or knowledge construct areas, median validity coefficients for training criteria range from .10 to .42 (the median of these median values is .27). For Job Proficiency criteria, median values range from .03 to .41, with the median of the medians at .10. Ghiselli (1966) reported similar differences between validity coefficients computed for training criteria and those computed using job performance measures.

Second, throughout the discussion of this final set of summary data, we focused exclusively on validity data for Training and Job Proficiency criteria. Coefficients computed using Educational criteria indicate that virtually all predictors are useful in predicting course grades or instructor rankings (with the exception of one cell, all median values are equal to or greater than .15). Median values computed across all job types for each predictor range from .16 to .38. Note that no data were located for Memory and Science Knowledge in this criterion category.

Finally, only a few correlations computed using Adjustment criteria were identified and reported in this summary. In our literature search, we emphasized research reporting validity data computed from training or job performance criteria. Thus, it not surprising that so few correlations between predictors and adjustment criteria were located. The small number of validities that we did locate confirmed our initial expectations. That is, cognitive ability measures are less effective at predicting adjustment outcomes than at predicting training or job performance outcomes. Overall, the median values for this criterion are near zero, with the exception of PS&A (.20) and Mechanical Aptitude (.23); both of these values were obtained from a single study.

SECTION VI

SUMMARY AND CONCLUSIONS

SUMMARY OF MAJOR FINDINGS

Researchers have been investigating the composition of intellectual abilities for nearly a century. Although early studies focused on methods for assessing general intelligence, the development of a statistical technique, factor analysis, led to systematic examination of the makeup of intelligence. Spearman, for example, postulated the existence of a single innate ability, g. According to his theory all specific abilities were learned, rather than innate. Thurstone, on the other hand, proposed that intelligence was composed of several distinct abilities. Results from his research indicated that at least seven primary mental abilities could be isolated. Guilford, at the extreme, suggested in his Structure-of-Intellect model that well over 120 separate ability factors can be identified using a matrix of content, operations, and products.

Although numerous researchers have formulated cognitive ability taxonomies, very few of these taxonomies have actually been implemented for practical applications. Thurstone's Primary Mental Ability battery of tests represents an example of one taxonomy that has actually been used in applied settings. Other cognitive ability batteries have been constructed for practical application in educational or work settings. Inspection of four of the most widely used batteries revealed that paper-and-pencil measures of cognitive ability are highly reliable (e.g., internal consistency and test-retest) and provide useful information about potential for success in educational and work settings.

Based on the cognitive abilities assessed in the four widely used test batteries and on two lines of extensive research into the abilities that comprise intelligence (i.e., Guilford's Structure-of-Intellect model and factor analysis data reported by researchers at Educational Testing Service), we constructed a cognitive taxonomy that contains nine ability factors: (1) Verbal, (2) Number Facility, (3) Spatial abilities, (4) Reasoning, (5) Memory, (6) Fluency, (7) Perception, (8) Perceptual Speed and Accuracy, and (9) Mechanical Aptitude. For seven of these ability factors, subfactors were identified and defined.

Although the notion of using measures of intelligence to make selection decisions in a work setting appeared during World War I when Yerkes was tasked with developing a measure to identify recruits unfit for military duty, it was not until later that researchers designed and administered a battery of cognitive ability tests to assist with vocational decisions. During the Depression, researchers at the Employment Stability Research Institute demonstrated that a battery of measures assessing a variety of personal characteristics could be used to make decisions related to individual vocational training needs. Also during this period, researchers began exploring the relationship between performance on cognitive ability tests and measures of job performance.

It was during World War II, however, that much of our knowledge about predictor measure and criterion measure development was provided. In particular, literally hundreds of cognitive ability tests were constructed and their validity for predicting training or job performance outcomes assessed. Many of these tests were used to select and classify recruits into different military occupations.

Following the war, similar test development and validation procedures were used to construct selection and classification devices for all military branches. Several test batteries have been constructed and used over the years in all of the military services; the current battery, the ASVAB, is used DOD-wide to select and classify recruits into occupational specialties. The battery contains ten subtests, seven measuring cognitive ability factors and three measuring knowledge in technical areas.

Coinciding with the development of measures of intelligence and specific cognitive abilities was the concern about possible bias in testing. Research in this area has proceeded along several avenues. Initially, mean test scores for different racial and ethnic subgroups were compared. Results from these research activities indicate that, indeed, mean scores for the majority and minority racial/ethnic subgroups differ and these differences are fairly consistent across several types of cognitive abilities. Mean scores for males and females may also differ, but the level of male-female test score differences varies according to the cognitive ability of interest. The question about why these differences appear for gender and racial subgroups remains unanswered.

Another avenue of test bias research has focused on correlations between cognitive ability measures and measures of educational, training, or job performance outcomes. In general, results from this line of research indicate that only on rare occasions do validities computed for different racial or ethnic subgroups differ significantly. The same is true of validities computed for male and female subgroups.

Closer inspection of validities computed for different subgroups, indicates that differences between components of the regression equation, computed separately for minority and non-minority subgroups, may be statistically significant. Most frequently, the intercepts are significantly different. In these situations, bias in test score interpretation may occur if a common regression equation is used. Although only limited data were available for the period covered by the literature review, evidence for differences between males and females suggests that components of the regression equation seldom result in bias in interpreting test scores.

In sum, our plan for identifying cognitive ability measures to supplement the ASVAB takes into account test bias issues and evidence documenting mean score differences between gender and racial or ethnic subgroups. Test construction and evaluation activities and validation procedures recommended by the Federal government serve to guide current project research activities.

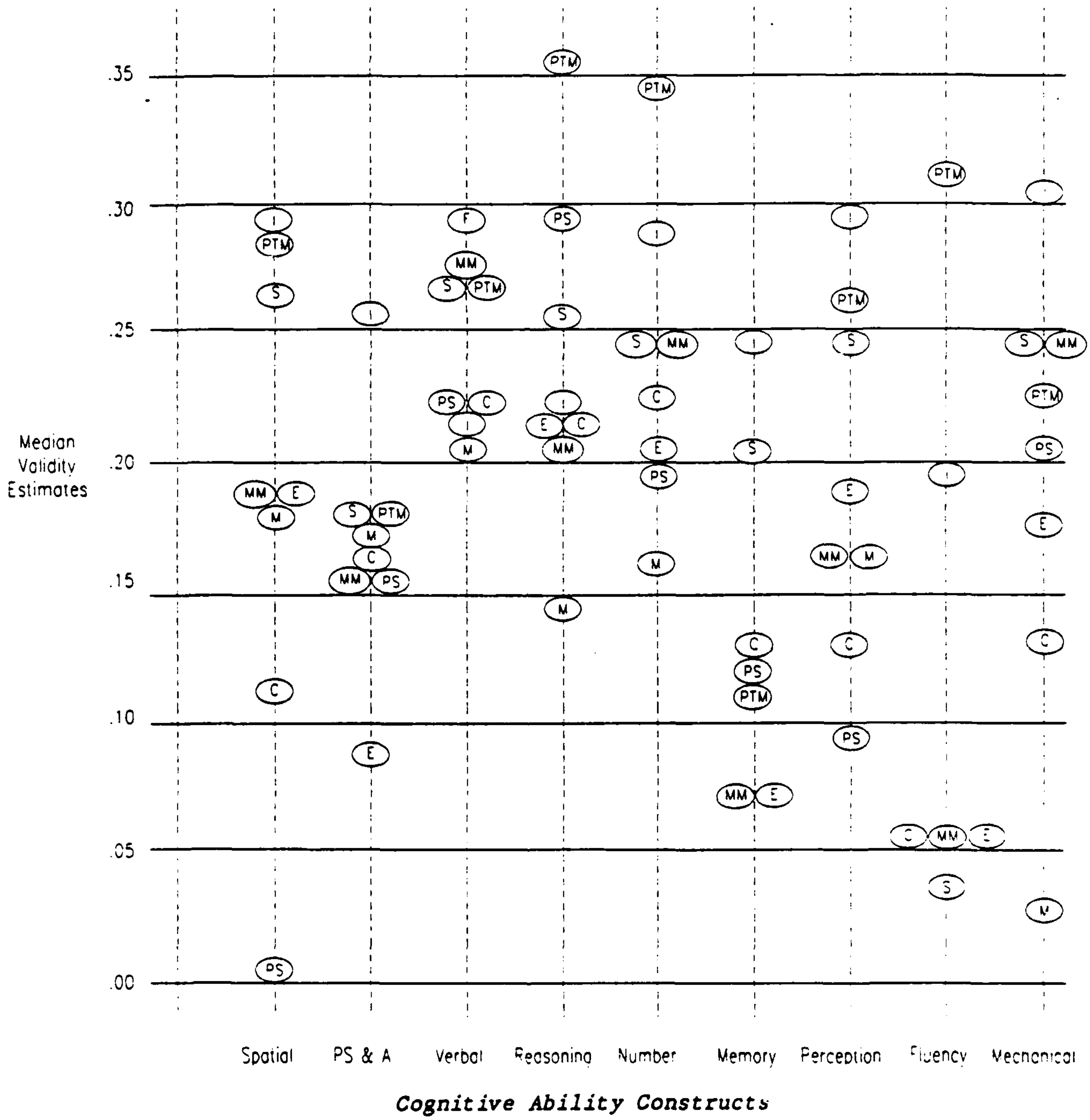
EVALUATION OF COGNITIVE ABILITY CONSTRUCTS

Identification of measures to supplement the ASVAB poses issues unique to the cognitive ability domain. That is, the current battery, as indicated previously, contains several cognitive ability measures. Thus, the first important element to consider, before identifying cognitive ability constructs for inclusion in an experimental battery of tests to supplement the ASVAB, is the content of the battery itself. Cognitive ability tests included in the ASVAB are (1) Word Knowledge, (2) Paragraph Comprehension, (3) Number Operations, (4) Mathematics Knowledge, (5) Arithmetic Reasoning, (6) Coding Speed, and (7) Mechanical Comprehension. According to results from a factor analysis of ASVAB subtest scores, the battery measures four ability areas (Kass et al., 1982). The first, verbal ability, is measured by Word Knowledge, Paragraph Comprehension, and General Science. The second factor, speeded performance, is measured by Coding Speed and Number Operations. Arithmetic Reasoning and Mathematics Knowledge combine to form a quantitative factor. A technical knowledge factor is formed from scores on Mechanical Knowledge, Electronics Information, and Auto/Shop Information.

A second important consideration involves the validity evidence summarized in the preceding section. Those data are condensed even more in Figures 3 and 4. In Figure 3, median validity coefficients are summarized by cognitive ability construct, and within each construct the median value is provided for each job type. In Figure 4, median validity coefficients are summarized by job type. Note that both figures present median uncorrected validity coefficients; corrected values and median values reported in Ghiselli's summary are <u>not</u> included in these figures. Median validity estimates recorded in these graphs are based on the Overall median computed in each Job Type and Predictor cell appearing in Table 45. We refer to these data as we evaluate each of the nine cognitive predictor constructs.

A final consideration in evaluating the constructs involves target Army MOS. In the early stages of Project A, staff identified 19 MOS that are representative of the nearly 300 occupational specialties for entry level personnel. During the time that we evaluated the cognitive ability constructs, project staff also conducted field site visits to observe recruits performing on the job. These job observations provided us with valuable information about job requirements and duties for many of the target MOS, such as tank crew members, cannon crewmen, MANPADS (Manned Personnel Air Defense Systems) personnel, military police, light wheel vehicle repairmen, radio and teletype operators, administrative specialists, and medical specialists. Evaluations of the cognitive ability constructs, then, were aided by the information gleaned from these job observations.

On the following pages, we evaluate the nine cognitive ability constructs to determine whether or not each might add unique variance to ASVAB selection and classification predictor equations. These evaluations are based in large part on the three factors listed above: (a) content of the ASVAB; (b) information gleaned from job observations; and (c) median validity coefficients obtained from the literature. For item (c), constructs with median validity coefficients equal to or above .15 are considered to be potentially useful for selection and classification purposes.



Job Type Codes: PTM = Professional, Technical, and Managerial, C = Clerical, PS = Protective Services, S = Service, MM = Mechanical and Structural Maintenance, E = Electronics, I = Industrial, and M = Miscellaneous

Figure 3. Graphic Display of Median Validities by Job Type for Nine Cognitive Ability Constructs

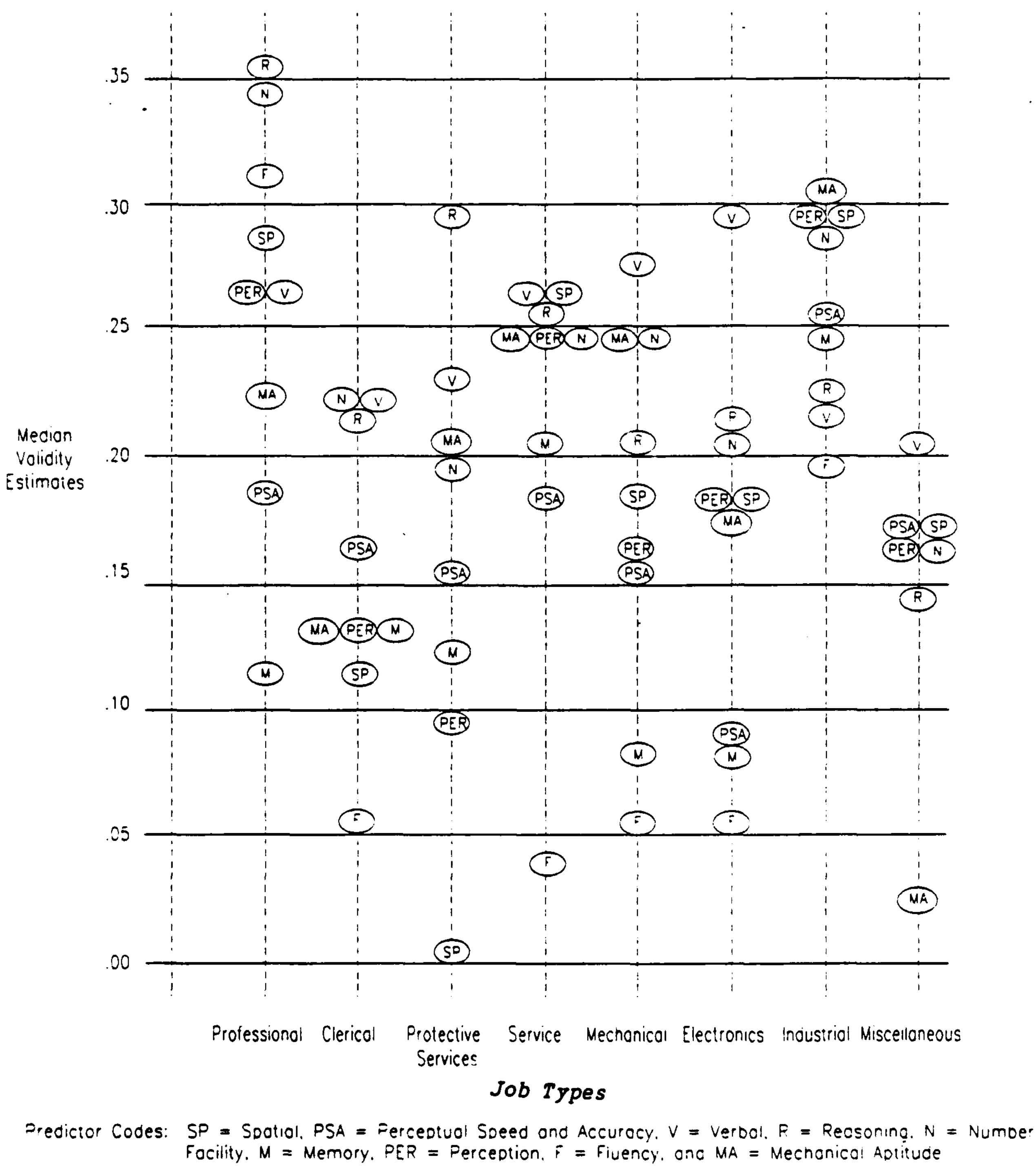


Figure 4. Graphic Display of Median Validities by Predictor for Eight Job Types

Using the three factors, each of the nine cognitive ability constructs is evaluated as having low, moderate, or high-priority development status. For those constructs with high-priority status, we examine specific measures that may be used to supplement information supplied by the ASVAB.

Spatial Ability

This construct involves the ability to visualize or rotate objects and figures in space. It is clear from the description above that the ASVAB contains no measures of spatial ability. According to the median validity estimates in Figure 3, spatial ability measures predict training and job performance outcomes for six of the eight job types included in the graph. Data in Figure 4 suggest that it is one of the best predictors for Service and Industrial occupations. Finally, observations of Army personnel performing on the job indicate that measures of spatial ability are potentially useful predictors of success on the job. For example, infantrymen, tank and cannon crew members, and MANPADS personnel are required to use maps to determine location in the field and to determine and maintain direction and orientation by using features in the environment. Thus, the spatial construct was assigned high priority for test development activities.

From the description of this construct provided in Table 7, it is clear that several types of measures may be constructed to assess spatial ability. Visualization tasks involve visually manipulating or transforming components of a figure to see how the components would appear under altered conditions. This ability is required for jobs that involve construction activities, mechanical maintenance, and so on.

Spatial rotation involves the ability to identify a two- or three-dimensional figure when seen at different angular rotations. Such abilities are required in Army MOS that involve identifying enemy vehicles or aircraft from different perspectives or directions. As indicated in Table 7, measures of two- and three-dimensional rotation are viewed as different abilities. Recall that, in our review of subgroup differences, males and females differ the most on measures of three-dimensional rotation. Thus, measures of two-dimensional rotation appear the most appropriate for development purposes.

Spatial scanning involves the ability to visually survey a complex field to find a particular configuration representing a pathway through a field. This ability is useful in jobs that involve electrical and electronics operations and using maps and diagrams.

A final spatial ability that surfaced in our review of the Army Air Forces research is spatial orientation. This involves the ability to maintain one's bearing with respect to points on a compass and to maintain or determine location relative to landmarks in the field. As noted above, this type of ability is required in many combat positions, such as infantrymen and MANPADS personnel.

Perceptual Speed and Accuracy

This construct represents the ability to perceive visual information quickly and accurately and to perform simple processing tasks with that information (e.g., make comparisons). From the summary data appearing in Figure 3, it appears that measures of this construct yield moderate validities for seven of the eight job types.

As noted above, one of the factors measured by the ASVAB is speeded performance, which includes both Coding Speed and Number Operations subtests. Although this construct appears to be adequately measured in the current selection battery, one concern with the subtests involves test length. That is, because both subtests are very short (7 minutes and 3 minutes, respectively), error may be introduced into scores if test administration is not accurately timed. Thus, more precise means of recording test responses may be desirable for this construct. Because it appears to be fairly well covered by the ASVAB, however, this construct was assigned only a moderate priority rating.

Verbal Ability

This construct represents the ability to understand the English language. The two subcomponents of this construct are (a) verbal comprehension, or knowledge of the meaning of words, and (b) reading comprehension, or ability to read and understand written material. Median validity coefficients presented in Figure 3 indicate that measures of this construct are highly valid for all job types. The current military battery contains two subtests, Word Knowledge and Paragraph Comprehension, that measure both components of this construct. Because additional measures of verbal ability appear unnecessary, no priority rating was assigned to this construct.

Reasoning

This construct involves the ability to discover a rule or principle and apply it in solving a problem. According to the median validity coefficients provided in Figure 4, Reasoning is one of the better predictors of training and job performance outcomes for Professional/Technical/Managerial, Protective Services, and Electronics occupations. Data in Figure 3 indicate that measures of this construct yield moderate validities across all job types.

The current battery contains a subtest, Arithmetic Reasoning, that appears to measure one of the subcomponents of the Reasoning construct: word problems. Results from the factor analysis study noted earlier (Kass et al., 1982), however, indicate that this ASVAB subtest corresponds more closely to measures of quantitative abilities. Further, field observations revealed that this ability is important for success in many Army MOS, such as military police. Given these facts, Reasoning was assigned a high development priority status.

Table 7 defines five subcomponents for the Reasoning construct. Note that the analogical reasoning and figural reasoning subcomponents are actually part of the inductive reasoning subcomponent, so this construct may be assessed using three types of measures. Inductive reasoning involves the ability to form and apply hypotheses that fit a set of data; as we noted, this may be assessed using items that contain verbal analogies or that involve reasoning with figures. Deductive reasoning is the ability to use logic and judgment in drawing conclusions from available information. Measures of reasoning that include word problems involve the ability to select and organize relevant information for mathematical problems. Based on observations of Army MOS and on content of the ASVAB, measures of inductive and deductive reasoning appear to have the greatest potential for contributing unique variance to prediction equations.

Number Facility

This construct involves the ability to solve simple or complex mathematical problems. Median validity coefficients reported in Figure 4 indicate that measures of Number Facility represent some of the better predictors of training and job performance criteria for Professional/Technical/Managerial, Clerical, Service, and Mechanical Maintenance occupations. Data reported in Figure 3 indicate that measures of this construct yield moderate to high validities across all job types.

According to our taxonomy, Number Facility contains two subcomponents (see Table 7). Again, results from the factor analysis study indicate that ASVAB subtests, Mathematical Knowledge and Arithmetic Reasoning, measure quantitative abilities; this corresponds to the subcomponent, use of formulations and number problems. Another ASVAB subtest, Number Operations, would appear to measure the second subcomponent, numerical computation. Results from the factor analysis study, however, place this subtest along with Coding Speed, producing a speeded performance factor. The test contains 50 multiple-choice items that require examinees to add, subtract, multiply, and divide single-digit items (e.g., 2-1, 8+8, 15/3, and 4x6). It appears, then, that this test measures ability to perform very simple arithmetic tasks.

Because the subcomponent, number computation, appears to be missing from the ASVAB, we are interested in developing an experimental measure that contains more complex items than those found in the Number Operations test. For this reason we have assigned this construct a moderate priority rating. If administration time permits, we will develop a new measure of number facility. Basically, however, this construct appears to be fairly well covered by ASVAB subtests.

Memory

Measures of this construct involve the ability to recall previously learned information or concepts. From the calculations provided in Table 45, it is clear that measures of this construct are used relatively less often than other types of cognitive predictor constructs in both military and non-military settings. According to the median values in Figure 3, measures of this ability yield moderate validities for two of the eight occupations. Further, results from Ghiselli's review indicate that Memory tests may be useful predictors of training and job performance criteria for Clerical, Protective Services, Service, and Mechanical Maintenance occupations (see Tables 32, 34, 36, and 38).

At present, the ASVAB contains no measures of memory abilities. Information collected in field observations indicates that such abilities are important for success in MOS that require recruits to accurately recall the sequence or order in which tasks must be performed. This particular ability appears critical for a number of Army MOS, such as cannon crewman, tank crewman, medical specialist, and infantryman. Thus, we assigned this construct a moderate to high priority status.

Perception

This construct involves the ability to perceive a figure or form that is partially presented or that is embedded in another form. Again, the ASVAB contains no such measures. Data in Figure 3 indicate that measures of Perception yield moderate validities for six of the eight occupations. Results from Ghiselli's review suggest that these types of measures are useful in predicting training and job performance outcomes in five of the eight occupational groups (see Tables 30, 32, 34, 38, and 42).

Information gleaned from field observations indicates that this ability is important for success in many combat and combat support MOS. Recruits in these types of MOS are required to detect camouflaged enemy vehicles and personnel in field settings. Because this ability appears useful for many combat occupations, we assigned this construct a moderate to high priority status.

Definitions of the two Perception subcomponents are provided in Table 7. The first, flexibility of closure, involves the ability to "hold" a given percept or configuration in mind so as to disembed it from other well-defined or complex material. This particular ability corresponds very closely to the ability to detect enemy vehicles or personnel. The second subcomponent, speed of closure, involves the ability to identify objects or words, given partial or sketchy information.

Fluency

Fluency involves the ability to rapidly generate words or ideas related to target stimuli. This particular construct is not measured by any ASVAB subtest. As we reported in the previous section, very few studies employed measures of this construct. Results from those studies that did use such measures indicate that it may be useful for Professional/Technical/Managerial and Industrial occupations. Results from Ghiselli's review also suggest that this particular construct is seldom used to predict training or job performance outcomes in the eight occupational groups. Given the limited amount of data, we concluded that measures of fluency might be useful for predicting success in higher level positions (e.g., noncommissioned officer potential), rather than entry-level occupations. For this reason, we assigned this construct a low priority rating.

Mechanical Aptitude

Measures of this construct assess the ability to perceive and understand the relationships of physical forces and mechanical elements in a prescribed situation. As noted previously, the current military selection and classification battery contains a measure of this construct. Data summarized in Figure 3 indicate that Mechanical Aptitude measures yield moderate validities for six of the eight occupational groups.

Subgroup mean score differences for males and females, specifically those reported for the ASVAB subtest, Mechanical Comprehension, are fairly high relative to other cognitive ability constructs (see Tables 18 and 19). A review of similar measures of mechanical aptitude reveals that many of the items contain questions about parts and equipment potentially more familiar to males than females. Although a fairly low priority status was assigned to this construct, we considered developing mechanical aptitude items that would be equally familiar to males and females.

CONCLUSIONS

Predictor Constructs. Based on our evaluation of the nine predictor constructs, it is clear that several constructs in the classic psychometric literature remain untapped by the current selection and classification battery. Given that these constructs are likely to add unique variance to prediction equations, preliminary priority status ratings suggest that measures of the following constructs be developed:

1. Spatial abilities
2. Reasoning
3. Perception
4. Memory

An important consideration for test development activities is the time allotted for experimental test battery administration. This includes time required to administer all parts of the experimental battery--that is, cognitive, non-cognitive, and psychomotor measures. Given this factor, if time permits, development of measures for three additional cognitive ability constructs--Number Facility, Perceptual Speed and Accuracy, and Mechanical Aptitude--merits consideration.

REFERENCES

Adcock, C. J. (1972). [Review of the Flanagan Industrial Tests]. In O. K. Buros (Ed.), The seventh mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Albemarle Paper Company v. Moody, 10 FEP 1181. (1975).

American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Anastasi, A. (1937). Differential psychology: Individual and group differences in behavior. New York: Macmillan.

Anastasi, A. (1964). Fields of applied psychology. New York: McGraw-Hill.

Anastasi, A. (1976). Psychological testing (4th ed.). New York: Macmillan.

Anastasi, A. (1979). Fields of applied psychology (2nd ed.). New York: McGraw-Hill.

Anastasi, A. (1983). Evolving trait concepts. American Psychologist, 38, 175-184.

Anastasi, A., & Foley, J. D., Jr. (1949). Differential psychology. New York: Macmillan.

Anderson, M. (1951). Woman at work: Autobiography of Mary Anderson as told to Mary N. Winslow. Minneapolis: University of Minnesota Press.

Anderson, R. N. (1929). Measurement of clerical ability: A critical review of proposed tests. Personnel Journal, 8, 232-244.

Angelino, H., & Shedd, C. L. (1955). An initial report of a validation study on the Davis-Eells tests of general intelligence or problem-solving ability. Journal of Psychology, 40, 35-38.

Arvey, R.D. (1972). Some comments on culture fair tests. Personnel Psychology, 25, 433-448.

Arvey, R. D. (1979). Fairness in selecting employees. Reading, MA: Addison-Wesley Publishing Company.

Atkinson, R. L., Atkinson, R. C., & Hilgard, E. R. (1983). Introduction to psychology (8th ed.). San Diego: Harcourt Press.

Bakke v. Regents of the University of California, 17 FEP 1. (1978).

Bannatyne, A. (1975). The DAT career planning program. Journal of Learning Disabilities, 8, 70-72.

Bartlett, C. J., Bobko, P., & Pine, S. M. (1977). Single-group validity: Fallacy of the facts? Journal of Applied Psychology, 62, 155-157.

Bechtoldt, H. P. (1947). Factorial investigation of the perceptual speed factor. American Psychologist, 2, 304-305.

Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. Behavioral and Brain Sciences, 11, 169-232.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1973). Manual for the Differential Aptitude Test. New York: The Psychological Corporation.

Binet, A., & Henri, V. (1895). La psycholgie individuelle. Annee Psychologique, 2, 411-463.

Binet, A., & Simon, T. (1905). The development of intelligence in children. L'Annee Psychologique, 11, 163-191.

Bobko, P., & Bartlett, C. J. (1978). Subgroup validities: Differential definitions and differential prediction. Journal of Applied Psychology, 63, 12-14.

Boehm, V. E. (1972). Negro-white differences in validity of employment and training selection procedures: Summary of research evidence. Journal of Applied Psychology, 56, 33-39.

Boehm, V. E. (1977). Differential prediction: A methodological artifact? Journal of Applied Psychology, 62, 146-154.

Bouchard, T. J. (1978). The differential aptitude test. From O. K. Buros (Ed.), The eighth mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Bowman, M. L. (1989). Testing individual differences in ancient China. American Psychologist, 2, 304-305.

Bruce, M. (1940). Factors affecting intelligence test performance of whites and Negroes in the rural South. Archives of Psychology, 252.

Burt, C. L. (1949). The structure of the mind: A review of the results of factor analysis. British Journal of Educational Psychology, 19, 100-111, 176-199.

Carroll, J. B. (1972). Stalking the wayward factors: Review of J. P. Guilford and Ralph Hoepfner, The Analysis of Intelligence. Contemporary Psychology, 17, 321-324.

Carroll, J. B. (1982). The measurement of intelligence. In R. J. Sternberg (Ed.), Handbook of human intelligence (pp. 31-120). London: Cambridge University Press.

Cascio, W. F. (1978). Applied psychology in personnel management. Reston, VA: Reston Publishing Company.

Cattell, J. M., & Ferrand, L. (1896). Physical and mental measurements of the students of Columbia University. Psychological Review, 3, 618.

Clark, R. C. (1983). Minnesota survey: State of the art in science education. St. Paul: Minnesota State Department of Education.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Cole, N. S. (1973). Bias in selection. Journal of Educational Measurement, 10, 237-255.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., MacPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966a). Equality of educational opportunity. Washington, DC: U.S. Office of Education.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., MacPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966b). Equality of educational opportunity, Supplemental appendix, Section 9.10. Washington, DC: U.S. Office of Education.

Coleman, W., & Ward, A. W. (1955). A comparison of Davis-Eells and Kuhlman-Finch scores of children from high- and low-socioeconomic status. Journal of Educational Psychology, 46, 463-469.

Comer, J. P. (1980). School power. New York: The Free Press

Comer, J. P. (1986). The Yale-New Haven primary prevention project: A follow-up study. Journal of American Academy of Child Psychiatry, 24, 154-160.

Comer, J. P. (1987, March). New Haven's school-community connection. Educational Leadership. 44, 13-18.

Connecticut v. Teal, 102 S. Ct. 2525. (1982).

Csikszentmihalyi, M., & Getzels, J. W. (1970). Concern for discovery: An attitudinal component of creative production. Journal of Personality, 38, 91-105.

Daniels, N. (1976). IQ, intelligence and educability. Philosophical Forum, 6, 56-69.

Darlington, R. B. (1971). Another look at "culture fairness." Journal of Educational Measurement, 8, 71-82.

Davis, A., & Eells, K. (1953). Davis-Eells Test of General Intelligence or Problem Solving Ability. Chicago: World Book.

Dellas, M., & Gaier, E. L. (1970). Identification of creativity: The individual. Psychological Bulletin, 73, 55-73.

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

Department of Labor. (1970). General Aptitude Test Battery Manual, Sections II, IIa and III. Washington, DC: Employment and Training Administration, U.S. Employment Service.

Department of Labor. (1977). Dictionary of occupational titles (4th ed.). Washington, DC: Employment and Training Administration, Government Printing Office.

Department of Labor. (1980). General Aptitude Test Battery Manual, Sections II, IIa and III (rev. ed.). Washington, DC: Employment and Training Administration, Government Printing Office.

Diaz v. Pan American Airways, 422 F. 2d. 385. (1971).

Doering, Z. D., Eitelberg, M. J., & Sellman, W. S. (1982). Uniforms and jeans: A comparison of 1981 military recruits with the 1980 national youth population. Paper presented at the Annual Meeting of the American Psychological Association, Washington, DC.

Doyle, K. O. (1974). Theory and practice of ability testing in ancient Greece. Journal of the History of Behavioral Sciences, 10, 202-212.

Dreger, R. M., & Miller, K. S. (1960). Comparative psychological studies of Negroes and whites in the United States. Psychological Bulletin, 57, 361-402.

DuBois, P. H. (Ed.). (1947). The classification program. Army Air Force Aviation Psychology Program Reports, No. 2. Washington, DC: Government Printing Office.

DuBois, P. H. (1966). A test-dominated society: China, 1115 B.C. - 1905 A.D. In A. Anastasi (Ed.), Testing problems in perspective (pp. 29-36). Washington, DC: American Council on Education.

DuBois, P. H. (1970). A history of psychological testing. Boston: Allyn and Bacon.

Dunnette, M. D. (1966). Personnel selection and placement. Belmont, CA: Wadsworth.

Dunnette, M. D. (1976). Aptitudes, abilities, and skills. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 469-520). Chicago: Rand McNally College Publishing Company.

Dunnette, M. D., & Borman, W. C. (1979). Personnel selection and classification systems. Annual Reviews of Psychology, 30, 477-525.

Ebbinghaus, H. (1897). Uber eine neue methode zur prufung geistiger fahigkeiten und ihre andwendung bei schulkindern. Zeitschrift Psychologie, 13, 401-459.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. (1951). Intelligence and cultural differences. Chicago: University of Chicago Press.

Egbert, R. L., Meeland, T., Cline, V. B., Forgy, E. W., Spickler, M. W., & Brown, C. (1958). Fighter I: An analysis of combat fighters and nonfighters (HumRRO Technical Report 44). Washington, DC: Human Resources Research Office.

Einhorn, H. J., & Bass, A. R. (1971). Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 75, 261-269.

Eitelberg, M. J., Laurence, J. H., Waters, B. K., & Perelman, L. S. (1984). Screening for service: Aptitude and education criteria for military entry. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Installations and Logistics).

Ekstrom, R. B., French, J. W., & Harman, H. H. (1979). Cognitive factors: Their identification and replication. Multivariate Behavioral Research Monographs, No. 79-2.

Ekstrom, R. B., French, J. W., Harman, H. H., & Durman, D. (1976). Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Equal Employment Opportunity Commission. (1966, August). Guidelines on employee selection procedures (29 C.F.R. 1607). Washington, DC: Author.

Equal Employment Opportunity Commission. (1970). Guidelines on employee selection procedures. Federal Register, 35, 12333-12336.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Adoption by four agencies of uniform guidelines on employee selection procedures. Federal Register, 43, 38290-38315.

Ernest, J. (1976). Mathematics & sex. American Mathematical Monthly, 83, 595-612.

Eysenck, H. J. (1953). Uses and abuses of psychology. London: Pelican.

Eysenck, H. J. (1967). Intelligence assessment: A theoretical and experimental approach. British Journal of Educational Psychology, 37, 81-98.

Fennema, E. & Sherman, J. (1977). Sex related differences in mathematics achievement, spacial visualization, and sociocultural factors. American Educational Research Journal, 14, 51-71.

Fine, B. (1975). The stranglehold of the I.Q. New York: Doubleday.

Flanagan, J. C. (Ed.). (1947). The Aviation Psychology Program in the Army Air Force. Army Air Force Aviation Psychology Programs, No. 1. Washington, DC: Government Printing Office.

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-425.

Flanagan, J. C. (1965). Flanagan Industrial Tests Manual. Chicago: Science Research Associates.

Flanagan, J. C. (1975). Flanagan industrial tests: Examiner's manual. Chicago: Science Research Associates.

Fowler, W. L. (1957). A comparative analysis of pupil performance on conventional and culture-controlled mental tests. Yearbook of the National Council on Measurements Used in Education, 14, 8-19.

Fox, C. H. (1977). The effects of sex role socialization on mathematics participation and achievement. In J. Schoemaker (Ed.), Women and mathematics: Research perspectives for change. National Institute of Education Papers in Education and Work, No. 8. Washington, DC: Education and Work Group, The National Institute of Education, Department of Health, Education and Welfare.

Fox, C. H. (1982). The study of social processes that inhibit or enhance the development of competence and interest in mathematics among highly able young women. Report to the National Institute of Education.

Frederiksen, N. (1959). [Review of the Primary Mental Abilities Test.] In O. K. Buros (Ed.), The fifth mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Frederiksen, N., Carlson, S., & Ward, W. C. (1984). The place of social intelligence in a taxonomy of cognitive abilities. Intclligence, 8, 315-337.

French, J. W. (1951). The description of aptitude and achievement tests in terms of rotated factors. Psychometric Monographs, 5.

French, J. W. (1954). Kit of selected tests for reference aptitude and achievement factors. Princeton, NJ: Educational Testing Service.

French, J. W., Ekstrom, R. B., & Price, L. A. (1963). A kit of reference tests for cognitive factors. Princeton, NJ: Educational Testing Service.

Galton, F. (1869). Hereditary genius: An inquiry into its laws and consequences. New York: Appleton.

Getzels, J. W., & Jackson, P. W. (1962). Creativity and intelligence: Explorations with gifted students. New York: Wiley.

Ghiselli, E. E. (1966). The validity of occupational aptitude tests. New York: Wiley.

Gibson, J. J. (1947). Motion picture testing and research. Army Air Force Aviation Psychology Program Reports, No. 7. Washington, DC: Government Printing Office.

Goddard, H. H. (1912). The Kallikak family. New York: Macmillan.

Goslin, D. A. (1963). The search for ability: Standardized testing in perspective. New York: Wiley.

Griggs v. Duke Power, 3 FEP 175. (1971).

Guilford, J. P. (1967). The nature of human intelligence. New York: McGraw-Hill.

Guilford, J. P. (1981). Higher-order structure-of-intellect abilities. Multivariate Behavioral Research, 16, 411-435.

Guilford, J. P. (1982). Cognitive psychology's ambiguities: Some suggested remedies. Psychological Review, 89, 48-59.

Guilford, J. P., & Hoepfner, R. (1971). The analysis of intelligence. New York: McGraw-Hill.

Guilford, J. P., & Lacey, J. I. (Eds.). (1947). Printed classification tests. Army Air Force Aviation Psychology Research Program Reports, No. 5. Washington, DC: Government Printing Office.

Guttman, L. (1958). What lies ahead for factor analysis? Educational and Psychological Measurement, 18, 497-515.

Haggard, E. A. (1954). Social status and intelligence: An experimental study of certain cultural determinants of measured intelligence. Genetic Psychology Monographs, 49, 141-186.

Hanna, G. S. (1978). [Review of the Differential Aptitude Test.] In O. K. Buros (Ed.), The eighth mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Harman, H. H. (1976). Modern factor analysis. Chicago: University of Chicago Press.

Harrell, T. W., & Churchill, R. D. (1941). The classification of military personnel. Psychological Bulletin, 38, 331-353.

Harris, M. L., & Harris, C. W. (1971). A factor analytic interpretation strategy. Educational and Psychological Measurement, 31, 589-606.

Harris, M. J., Mitchell, F. D., & Schechter, S. J. (1984). The homefront: America during World War II. New York: G. P. Putnam's Sons.

Haven, E. W. (1971). Factors associated with the selection of advanced academic mathematics courses by girls in high school. Doctoral dissertation, University of Pennsylvania. Dissertation Abstracts International, 32, 1747A.

Haynes, J. R. (1970). Hierarchical analysis of factors in cognition. American Educational Research Journal, 7, 55-68.

Herman, D. O. (1978). [Review of the Flanagan Industrial Tests.] In O. K. Buros (Ed.). The seventh mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Hicks v. Crown Zellerbach, 2 FEP 1059. (1970).

Hocevar, D. (1980). Intelligence, divergent thinking and creativity. Intelligence, 4, 25-41.

Horn, J. L. (1967). On subjectivity in factor analysis. Educational and Psychological Measurement, 27, 811-820.

Horn, J. L. (1972). [Review of the Flanagan Industrial Tests.] In O. K. Buros (Ed.), The seventh mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Horn, J. L., & Knapp, J. R. (1973). On the subjective character of the empirical base of Guilford's Structure-of-intellect model. Psychological Bulletin, 80, 33-43.

Humphreys, L. G. (1973). Statistical definitions of test validity for minority groups. Journal of Applied Psychology, 58, 1-4.

Humphreys, L. G. (1986). Commentary. Journal of Vocational Behavior, 29, 421-437.

Hunter, J. E., Schmidt, F. L., & Rauchenberger, J. M. (1977). Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. Journal of Applied Psychology, 62, 245-260.

Jensen, A. R. (1980). Bias in mental testing. New York: The Free Press.

Kamp, J. D., & Hough, L. M. (1988). Utility of temperament for predicting job performance. In L. M. Hough (Ed.), Literature Review: Utility of Temperament, Biodata, and Interest & Assessment for Predicting Job Performance (ARI Research Note 88-02). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences. (AD A192 109)

Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1983). Factor structure of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10: 1981 applicant sample (ARI Technical Report 581). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A135 660)

Katz, I., & Greenbaum, C. (1963). Effects of anxiety, threat, and racial environment on task performance on Negro college students. Journal of Abnormal and Social Psychology, 66, 562-567.

Katzell, R. A., & Dyer, F. J. (1977). Differential validity revived. Journal of Applied Psychology, 62, 137-145.

Kelley, T. L. (1940). Comment on Wilson and Worchester's "Note on Factor Analysis." Psychology, 5, 117-120.

Klineberg, O. (1963). Negro-white differences in intelligence test performance: A new look at an old problem. American Psychologist, 18, 198-203.

Kurtz, A. K. (1959). [Review of the Primary Mental Abilities Test.] In O. K. Buros (Ed.), The fifth mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Levin, H. M. (1988). Commentary: Issues of agreement and contention in employment testing. Journal of Vocational Behavior, 33, 398-403.

Linn, R. L. (1978a). [Review of the Differential Aptitude Test.] In O. K. Buros (Ed.), The eighth mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Linn, R. L. (1978b). Single-group validity, differential validity and differential prediction. Journal of Applied Psychology, 63, 507-512.

Ludlow, H. G. (1956). Some recent research on the Davis-Eells game. School and Society, 84, 146-148.

Maccoby, E. E., & Jacklin, C. N. (1974). The psychology of sex differences. Stanford, CA: Stanford University Press.

MacKinney, A. C. (1978). [Review of the Flanagan Industrial Tests.] In O. K. Buros (Ed.), The eighth mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Massey, I. H., & Creagor, J. A. (1956). Validation of the Airman Classification Battery: 1949-53 (Report No. AFPTRC-TN-56-129). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center.

Matarazzo, J. D. (1972). Wechsler's measurement and appraisal of adult intelligence (5th ed.). Baltimore, MD: Williams and Wilkins.

McGurk, F. C. J. (1953a). On white and Negro test performance and socio-economic factors. Journal of Abnormal and Social Psychology, 48, 448-450.

McGurk, F. C. J. (1953b). Socioeconomic status and culturally-weighted test scores of Negro subjects. Journal of Applied Psychology, 37, 276-277.

McGurk, F. C. J. (1967). The culture hypothesis and psychological tests. In R. E. Kuttner (Ed.), Race and modern science (pp. 367-381). New York: Social Science Press.

McGurk, F. C. J. (1975). Race differences--twenty years later. Homo, 26, 219-239.

Myart v. Motorola, 110 Congressional Record 5662-64. (1964, March).

Nelson, R. J. (1955). The ten-year story of the industrial relations center. Bulletin, 15. Minneapolis: University of Minnesota Press.

Noll, V. H. (1958). Relation of scores on Davis-Eells Test of General Intelligence to social status, school achievement and other intelligence test results. American Psychologist, 13, 394.

O'Connor, J. (1943). Structured visualization. Boston: Human Engineering Laboratory.

Oltman, K., Raskin, E., & Witkin, H. (1971). Group Embedded Figures Test. Palo Alto, CA: Consulting Psychologists Press.

O'Sullivan, M. (1983). Recognizing facial expressions of emotion. Paper presented at the Annual Meeting of the American Psychological Association, Anaheim, CA.

O'Sullivan, M., & Guilford, J.P. (1976). Four factor tests of social intelligence (behavioral cognition), Manual of instructions and interpretations. Orange, CA: Sheridan Psychological Service.

Parsons, J. E., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. C., & Midgley, C. (1983). Expectancies, values, and academic behavior. In J. T. Spence (Ed.), Achievement and achievement motives: Psychological and sociological approaches. San Francisco, CA: W. H. Freeman.

Parsons, J. E., Adler, T. F., & Kaczala, C. (1982). Socialization of achievement, attitudes, and beliefs: Parental influence. Child Development, 53, 310-321.

Paterson, D. G., & Darley, J. G. (1936). Men, women, and jobs. Minneapolis: University of Minnesota Press.

Paterson, D. G., & Yoder, D. (1955). Industrial relations research: Ten years of progress. Minneapolis: University of Minnesota Press.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophy Magazine, Series 2(6), 559-572.

Quereshi, M. Y. (1972). [Review of the Differential Aptitude Test.] In O. K. Buros (Ed.), The seventh mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Ramey, C. T., Breyart,D. M., et al. (1988). Early intervention for high risk children, The California early intervention program. In R. H. Price, E. Cowen, et al. (Eds.), Fourteen ounces of prevention: A casebook for practitioners. Washington, DC: American Psychological Association.

Resnick, S. M. (1982). Psychological functioning in individuals with congenital adrenal hyperplasia: Early hormonal influences on cognition and personality. Unpublished doctoral dissertation. Minneapolis: University of Minnesota.

Ross, P. F. (1965). [Review of the Employee Aptitude Survey.] In O. K. Buros (Ed.), The sixth mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Ruch, F. L., & Ruch, W. W. (1963). Employee Aptitude Survey: Technical report. Los Angeles, CA: Psychological Services.

Ruch, F. L., & Ruch, W. W. (1980). Employee Aptitude Survey: Technical report (rev. ed.). Los Angeles, CA: Psychological Services.

Ruch, W. W. (1972, September). A re-analysis of published differential validity studies. Symposium paper presented at Annual Meeting of the American Psychological Association, Honolulu, Hawaii.

Schmidt, F. L. (1988). The problem of group dynamics in ability test scores in employment selection. Journal of Vocational Behavior, 33, 272-292.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. (1973). Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 58, 5-6.

Schmitt, N., Mellon, P. M., & Bylenga, C. (1978). Sex differences in validity for academic and employment criteria, and different types of predictors. Journal of Applied Psychology, 63, 145-150.

Schultz, R. S. (1936). Standardized tests and statistical procedures in selection of life insurance sales personnel. Journal of Applied Psychology, 29, 553-566.

Schutz, R. E. (1965). [Review of the Differential Aptitude Test.] In O. K. Buros (Ed.), The sixth mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Science Research Associates. (1965). Primary mental abilities technical report. Chicago: Science Research Associates.

Sevy, B. A. (1982). Sex-related differences in spatial ability: A meta analytic review of fifty years of research. Unpublished manuscript. Minneapolis: University of Minnesota.

Sevy, B. A. (1988). Adverse impact: Small differences have large effects. Unpublished manuscript. Minneapolis: University of Minnesota.

Shartle, C. L., & Dvorak, B. J. (1943). Occupational analysis activities in the War Manpower Commission. Psychological Bulletin, 40, 701-713.

Sherman, J. A. (1967). Problem of sex differences in space perception and aspects of intellectual functioning. Psychological Review, 74, 290-299.

Sherman, J. A. (1974). Field articulation, sex, spatial visualization, dependency, practice, laterality of the brain, and birth order. Perceptual and Motor Skills, 38, 1223-1235.

Singer, D. L., & Whiton, M. B. (1971). Ideational creativity and expressive aspects of human figure drawing in kindergarten-age children. Developmental Psychology, 4, 366-369.

Smith, P. C., & Kendall, L. M. (1965). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Spearman, C. (1904). General intelligence, objectively determined and measured. American Journal of Psychology, 15, 201-293.

Staff of the Psychological Branch. (1943). The Aviation Psychology Program of the Army Air Force. Psychological Bulletin, 40, 759-769.

Staff, Personnel Research Section. (1943a). Personnel research in the Army: I. Background and organization. Psychological Bulletin, 40, 129-135.

Staff, Personnel Research Section. (1943b). Personnel research in the Army. III. Some factors affecting research in the Army. Psychological Bulletin, 40, 271-278.

Staff, Personnel Research Section. (1945). The Army General Classification Test. Psychological Bulletin, 42, 760-768.

Staff, Personnel Research Section. (1947). The Army General Classification Test, with special reference to the construction and standardization of Forms 1a and 1b. Journal of Educational Psychology, 30, 385-420.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. Journal of Personality and Social Psychology, 41, 37-55.

Stevenson, R. A. (1931, November). The Minnesota unemployment research project. Bulletin of the ESRI, 1(1). Minneapolis: University of Minnesota Press.

Super, D. E., & Crites, J. O. (1962). Appraising vocational fitness by means of psychological tests. New York: Harper.

Tanser, H. A. (1939). The settlement of Negroes in Kent County, Ontario. Chatham, Ontario: Shepherd Publishing Company.

Taylor, C. W. (1947). A factorial study in fluency of writing. Psychometrika, 12, 239-262.

Taylor, E. K. (1965). [Review of the Employee Aptitude Survey.] In O. K. Buros (Ed.), The sixth mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Terkel, S. (1984). The good war: An oral history of World War II. New York: Pantheon.

Thomas, P. J. (1970). Validation of Form 7 of the USN Basic Test Battery in 108 class "A" schools (NPTRL Technical Bulletin STB 70-7). San Diego: Naval Personnel and Training Research Laboratory.

Thomas, P. J., & Thomas, E. D. (1965) School validation of the Basic Test Battery and of other classification tests 1962-1963 (Technical Bulletin STB 66-8). San Diego, CA: U.S. Naval Personnel Research Activity.

Thorndike, E. L., et al. (1921). Intelligence and its measurement: A symposium. Journal of Educational Psychology, 12, 123-147; 195-216; 271-275.

Thorndike, E. L., Bregman, E. O., & Cobb, M. V. (1927). The measurement of intelligence. New York: Teacher's College, Columbia University.

Thorndike, R. L. (Ed.). (1947). Research problems and techniques. Army Air Force Aviation Psychology Program, No. 3. Washington, DC: Government Printing Office.

Thorndike, R. L. (1971). Concepts of culture fairness. Journal of Educational Measurement, 8, 63-70.

Thurstone, L. L. (1938a). Primary mental abilities. Psychological Monographs, No. 1. Chicago: University of Chicago Press.

Thurstone, L. L. (1938b). The perceptual factor. Psychometrika, 3, 1-17.

Thurstone, L. L. (1944). A factorial study of perception. Chicago: University of Chicago Press.

Thurstone, L. L., & Jeffrey, T. E. (1979). Flags: A test of space thinking. Chicago: University of Chicago Press.

Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. American Psychologist, 3, 54-56.

Tuddenham, R. D. (1963). The nature and measurement of intelligence. In L. Postman (Ed.), Psychology in the making. New York: Alfred A. Knopf.

Tyler, L. G. (1965). The psychology of human differences. New York: Appleton-Century-Crofts.

Uhlaner, J. E. (1952). Development of Armed Forces Qualification Test and predecessor Army screening tests, 1946-1950. TRR 976 (Army Project Number 29531100). Washington, DC: Personnel Research Branch.

United States v. Georgia Power Company, 474 F. 2d, 906. (1973).

United Steelworkers of America v. Weber, 443 U.S. 193. (1979).

Vernon, P. E. (1961). The structure of human abilities (2nd ed.). London: Methuen.

Vernon, P. E. (1964). The structure of human abilities (2nd ed. rev.). London: Methuen.

Vernon, P. E. (1979). Intelligence: Heredity and environment. San Francisco: W. H. Freeman and Company.

Viteles, M. S. (1929). The human factor in substation operation: Specifications and tests for substation operators. Personnel Journal, 8, 81-113.

Vitola, B. M., Mullins, C. J., & Croll, P. R. (1973, July). Validity of Armed Services Vocational Aptitude Battery, Form 1, to predict technical school success (AFHRL-TR-73-7). Lackland Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Wallace, S. R. (1959). [Review of the Employee Aptitude Survey.] In O. K. Buros (Ed.), The fifth mental measurements yearbook. Highland Park, NJ: Gryphon Press.

Wallach, M. A., & Wing, C. S. (1969). The talented student: A validation of the creativity intelligence distinction. New York: Rineholt and Winston.

Washington v. Davis, 12 FEP 1415. (1976).

Willerman, L. (1979). The psychology of individual and group differences. San Francisco: W. H. Freeman and Company.

Williams, R. L. (1970). Black pride, academic relevance, and individual achievement. The Counseling Psychologist, 2, 18-22.

Williams, R. L. (1972). The BITCH test (Black Intelligence Test of Cultural Homogeneity). St. Louis, MO: Black Studies Program, Washington University.

Wilson, J. R., & Vandenberg, S. G. (1978). Sex differences in cognition: Evidence from the Hawaii family study. In T. E. McGill, D. A. Dewsbury, & B. D. Sachs (Eds.), Sex and behavior (pp. 317-333). New York: Plenum Press.

Wissler, C. (1901). The correlation of mental and physical traits. Psychological Monographs, 3(16).

Witkin, H. A., Lewis, H. B., Hertzman, M., Machover, K., Meissner, P. B., & Warner, S. (1954). Personality through perception: An experimental and clinical study. New York: Harper.

Woodrow, H. (1939). The common factors in fifty-two mental tests. Psychometrika, 4, 99-108.

Yerkes, R. M. (1921). Psychological examining in the United States Army (Vol. XV). Washington, DC: Memoirs of the National Academy of Sciences.

APPENDIX A

Description of Measures Included in the Army Air Force Classification Battery (December 1943)

APPENDIX A

Classification Battery Tests - December 1943

Test: Reading Comprehension (CI 616G)

Construct Measured: Verbal Ability

Description: The test contains six paragraphs with four to six questions about each paragraph. According to the authors, two paragraphs were targeted toward pilots, two toward bombardiers, and two toward navigators. The test contains a total of 30 items with a 30 minute time limit.

Test: Spatial Orientation I and II (CP 501B & CP 503B)

Construct Measured: Orientation

Description: Part I: Subjects are presented with a large aerial photograph along with six smaller photographs which are part of the larger photograph. The task is to match the small photographs with lettered sections of the large photograph. The test contains nine large aerial photographs with 49 scored items with a 5-minute time limit.

Part II. Subjects are presented with a standard aviation map sectioned off into 12 squares lettered A through M. The task is to match each square with a smaller aerial photograph presented below it. Subjects are presented with 13 aerial maps and must respond to 50 scored items with an 18-minute time limit.

Test: Dial and Table Reading (CP 622A and CP 621A)

Construct Measured: Perceptual Speed and Accuracy

Description: Part 1. Dial Reading: Subjects are presented with seven dials along with items indicating which dials are to be read. After identifying the appropriate dial, the subject just read it correctly and select the response that most closely matches the value indicated on the dial. The test contains 57 items with a 9-minute time limit.

Part 2. Table Reading: Subjects are asked to locate values given in a large table. A second part of this test provides subjects with four tables containing information related to flight of an airplane. Values are given for air speed, angle of wind and velocity of the wind. For each item, then, subjects must use the values to determine the drift correction or ground speed. Section I of Part 2 contains 43 items and a 4-minute time limit; Section II contains 43 items and a 7-minute time limit.

Test: Mechanical Principles (CI 903A)

Construct Measured: Mechanical Aptitude

Description: Subjects are presented a pictorial display of some activity and are asked to select the response that most accurately describes the action portrayed. The test contains 30 items with a 15-minute time limit.

Test: Arithmetic Reasoning (CI 206B)

Construct Measured: Reasoning

The test contains 30 problems that can be solved with minimal formal mathematical training. All items are formulated in aviation terms. Subjects are given 35 minutes to complete this test.

Test: Instrument Comprehension I and II (CI 615A CI 616A)

Construct Measured: Spatial

Description: Part I: Subjects are shown drawings of six instruments, altimeter, compass, airspeed, artificial horizon, rate-of-climb dial, and turnbank indicator. Subjects must select the correct written description from among the five presented. This part contains 15 items with a 12-minute time limit.

Part II: Subjects are presented with drawings of two instruments, compass and artificial horizon followed by five photographs each showing an airplane in a different position. Subjects must choose the picture that agrees most closely with the two instrument readings. This part contains 60 items with a fifteen minute time limit.

Test: Technical Vocabulary Pilot and Navigator (CE 505C)

Construct Measured: General Information

Description: The test contains three parts, each part is targeted toward one of the three aircrew positions: pilot, bombardier, or navigator. The 40 pilot items deal with planes, plane identification, and flying technique. The 40 navigator items deal with astronomy, instruments, and maps. The 20 bombardier items relate to guns, bomb sites, trajectories, etc. All items present a definitional statement completed by one of five response alternatives. Subjects are given 12 minutes to complete each part.

Test: Mathematics (CI 702E)

Construct Measured: Mathematics Ability

Description: This test is designed to measure ability and achievement in advanced arithmetic, algebra, and trigonometry. Subjects are asked to complete 30 items. (Time limit not reported.)

Test: Numerical Operations Front and Back (CI 702B)

Construct Measured: Numerical Facility

Description: On the first page subjects are presented with 100 addition and multiplication items along with answers to each. The task is to indicate whether each answer is correct (c) or wrong (w). The second page contains 80 subtraction and division items. The task here is to select the correct response from among four alternatives. Subjects are given 5 minutes to complete the first page and 5 minutes to complete the second page.

Test: Speed of Identification (CP 610A)

Construct Measured: Speed of Perceptual Detail

Description: Subjects are presented with four planes to the left of the page. To the right are five planes presented in different, rotated positions. The task involves matching the planes on the right with one of the four planes on the left; one plane does not match. The test includes 12 different plane groups with four items per group. Subjects are given 4 minutes to complete the 48 items.

Test: Biographical Data, Pilot and Navigator (CE 602D)

Construct Measured: Interests, Attitudes and Background

Description: Subjects are asked to provide information about home and personal history (20 items), interest in school subjects (10 items), interest in various activities (30 items), proficiency in sports (12 items), previous employment and occupational experience (9 items), military experience (10 items), preference for military and aircrew position (21 items), and degree of agreement with controversial statements (34 items). Contains a total of 65 items demonstrating empirical validity for pilot or navigator prediction. Subjects are given 25 minutes to complete this measure.

APPENDIX B

References for Validities Summarized in the Military and Non-Military Tables in Section V

REFERENCES FOR MILITARY AND NON-MILITARY VALIDITY TABLES

Bemis, S. E. (1968). Occupational validity of the GATB. Journal of Applied Psychology, 52, 240-241.

Bennett, G. K. (1969). Bennett Mechanical Comprehension Test Manual. New York: The Psychological Corporation.

Bennett, G. K., & Gelink, M. (1972). Short Employment Tests Manual. New York: The Psychological Corporation.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1973). Manual for the Differential Aptitude Test. New York: The Psychological Corporation.

Boone, J. O. (1979, September). Toward the development of a new selection battery for air traffic control specialists. Washington DC: Federal Aviation Administration, Office of Aviation Medicine.

Borich, G. D., & Bauman, P. M. (1972). Convergent and discriminant validation of the French and Guilford-Zimmerman spatial orientation and spatial visualization. Educational and Psychological Measurement, 32, 1029-33.

Campbell, C. H., & Black, B. A. (1982). Predicting trainability of M1 crewmen [TR-FR-MTRD (KY)-8-7]. Fort Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences.

Cory, C. H. (1976). An evaluation of computerized tests as predictors of job performance: II. Differential validity for global and job element criteria. (NPRDC-TR-76-28). San Diego: Navy Personnel Research and Development Center.

Cory, C. H. (1982). Assignment of general detail personnel in the Navy: Fleet follow-up of personnel appraised in a technical classification assessment center pilot study. (NPRDC-TN-82-23). San Diego: Navy Personnel Research Activity.

Curtis, E. W. (1968). Aptitudes, A-school grades, and subsequent performance in the CTR rating (USNPRA Technical Bulletin STB 69-3). San Diego: U.S. Naval Personnel Research Activity.

Curtis, E. W. (1971). Prediction of enlisted performance: I. Relationship among aptitude tests, Navy school grades, the report of enlisted performance evaluation, and advancement examinations (Technical Bulletin STB 71-10). San Diego: Naval Personnel and Training Research Laboratory.

Department of Labor. (1970). Manual for the USES General Aptitude Test Battery. Washington, DC: Manpower Administration.

Department of Labor. (1980). Manual for the USES General Aptitude Test Battery (rev. ed.). Washington, DC: Manpower Administration.

Eaton, N. K. (1978). Predicting tank gunnery performance (ARI-RM-78-6). Fort Knox, KY: Army Research Institute Field Unit.

Eaton, N. K., Bessemer, D. W., & Kristiansen, D. M. (1979). Tank crew position assignment (TR-391). Fort Knox, KY: Army Research Institute Field Unit.

Egbert, R. L., Meeland, T., Cline, V. B., Forgy, E. W., Spickler, M. W., & Brown, C. (1958, March). Fighter I: An analysis of combat fighters and nonfighters (HumRRO Technical Report 44). Washington, DC: Human Resources Research Office.

Field, H. S., Bayley, G. A., & Bayley, S. M. (1977). Employment test validation for minority and non-minority production workers. Personnel Psychology, I, 37-46.

Flanagan, J. C. (1965). Flanagan Industrial Tests Manual. Chicago: Science Research Associates.

Gael, S., Grant, D. L., & Ritchie, R. J. (1975). Employment test validation for minority and non-minority telephone operators. Journal of Applied Psychology, 60, 411-419.

Ghiselli, E. E. (1966). The validity of occupational aptitude tests. New York: Wiley.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Greenstein, R. B., & Hughes, R. G. (1977). The development of discriminators for predicting success in armor crew positions (ARI-RM-77-27). Fort Knox, KY: Army Research Institute Field Unit.

Guilford, J. P., & Zimmerman, W. S. (1956). The Guilford-Zimmerman aptitude survey: A manual of instructions and interpretations. Beverly Hills, CA: Sheridan Supply Company.

Helme, W.H. & White, R.K. (1958). Prediction of Job Performance in guided missile crew specialties. PRB-TR-89, Army Project No. 29560000.

Hitch, G. J. (1978). The numerical abilities of industrial trainee apprentices. Journal of Occupational Psychology, 51, 163-176.

Hoiberg, A., & Pugh, W. G. (1978). Predicting Navy effectiveness: Expectations, motivation, personality, aptitude, and background variables. Personnel Psychology, 31, 842-852.

Hunter, D. R. (1975). Development of an enlisted psychomotor/perceptual test battery: 1949 (AFHRL-TR-75-60). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Hunter, D. R., & Thompson, N.A. (1978). Pilots selection system development (AFHRL-TR-78-33). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Jensen, H. E., & Valentine, L. D., Jr. (1976). Validation of ASVAB-2 against civilian vocational-technical high school criteria (AFHRL-TR-76-16). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Johnson, C. D., Burke, L. K., Loeffler, J. C., & Drucker, A. J. (1955). Prediction of the combat proficiency of infantrymen (PRB-TR-1093). Washington, DC: The Adjutant General's Office, Personnel Research Branch.

Laney, A. R. (1951). Validity of employment tests for gas-appliance service personnel. Personnel Psychology, 4, 199-208.

Mackie, R. R., Ridihalgh, R. R., & Schultz, T. E. (1981). New criteria for the selection and evaluation of sonar technicians, Navy Personnel Research and Development Center, (NPRDC-TR-81-13). Goleta, CA: Human Factors Research.

Maier, M. H., & Grafton, F. C. (1981). Aptitude composites for ASVAB 8, 9, and 10. Fort Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences (Research Report 1308).

Martinek, H., Sadacca, R., & Burke, L. K. (1965). Development of a selection battery for Army image interpreters (USAPRO-TR-1148). Alexandria, VA: Defense Technical Information Center.

Massey, I. H., & Creagor, J. A. (1956). Validation of the Airman Classification Battery: 1949-1953 (AFPTRC-TN-56-129). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center.

Mathews, J. J. (1977). Analysis aptitude test for selection of airmen for the radio communications analysis specialist course: Development and validation (AFHRL-TR-77-74). Air Force Human Resources Laboratory, Brooks Air Force Base, TX: Air Force Systems Command.

Mathews, J. J., & Jensen, H. E. (1977, October). Screening test battery for dental laboratory specialist course: Development and validation. Brooks Air Force Base, TX: Personnel Research Division.

Mathews, J. J., & Roach, B. W. (1983). Reading abilities tests: development and norming for Air Force use (AFHRL11-TR-82-26). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Michael, W., Zimmerman, W. S., & Guilford, J. P. (1951). An investigation of the nature of the spatial relations and visualization factors in two high school samples. Educational and Psychological Measurement, 11, 561-577.

Osburn, H. G., Sheer, D. E., Elliott, J. M., Mullins, C. J. (1964). Validation of a carefulness test battery against training school criteria (PRL-TDR-64-7). Lackland Air Force Base, TX: Personnel Research Laboratory.

Price, L., & Eliot, J. (1975). Convergent and discriminant validities of two sets of measures of spatial orientation and visualization. Educational and Psychological Measurement, 35, 975-97.

Ronan, W. W. (1964). Evaluation of skilled trades performance predictors. Educational and Psychological Measurement, 24, 601-607.

Ruch, F. L., & Ruch, W. W. (1980). Employee Aptitude Survey: Technical Report. Los Angeles: Psychological Services.

Schmidt, B. J. (1975). Prediction of success in clerical occupations from ability test scores. Vocational Guidance Quarterly, 24, 68-72.

Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job sample vs. paper-and-pencil trade and technical tests: Adverse impact and employee attitudes. Personnel Psychology, 30, 187-197.

Science Research Associates. (1962). Primary Mental Abilities Technical Report. Chicago: Science Research Associates.

Science Research Associates. (1974). SRA Test Validation Studies. Chicago: Science Research Associates.

Siegal, A. I., & Bergman, B. A. (1972, June). Non-verbal and culture fair performance prediction procedures. I. Background, test development, and initial results. Wayne, PA: Applied Psychological Services.

Sparrow, J., Patrick, J., Speergeon, P., & Barwell, F. (1982). The use of job component analysis and related aptitudes in personnel selection. Journal of Occupational Psychology, 55, 157-164.

Suddick, D. E., & Bower, C. L. (1982). Longitudinal study of reading scores of the Stanford Achievement Test. Perceptual and Motor SKills, 54, 369-370.

Talbert, T. L., Carroll, K. I., & Ronan, W. W. (1976). Measuring clerical job performance. Personnel Journal, 55, 573-575.

Thomas, P. J. (1970). Validation of Form 7 of the USN Basic Test Battery in 108 Class "A" Schools (NPTRL Technical Bulletin STB 70-7). San Diego: Naval Personnel and Training Research Laboratory.

Thomas, P. J., & Thomas, E. D. (1965). School validation of the Basic Test Battery and of other classification tests 1962-1963 (Technical Bulletin STB66-8). San Diego: U.S. Naval Personnel Research Activity.

Valentine, L. D., Jr. (1977). Navigator-observer selection research: Development of new Air Force qualifying test navigator-technical composite (AFHRL-TR-77-36). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Wesman, A. G., & Doppelt, J. E. (1969). Personnel Tests for Industry Manual. New York: The Psychological Corporation.

Wilbourn, J. M., & Guinn, N. (1973). Feasibility of using special measures in the classification and assignment of lower mental ability airmen (AFHRL-TR-73-31). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Wilbourn, J. M., Guinn, N., & Leisey, S. A. (1976). Validation of non-verbal measures for selection and classification of enlisted personnel (AFHRL-TR-76-72). Lackland Air Force Base, TX: Personnel Research Division.

Wood, D. A. (1980). Investigation of possible sex differences in the prediction of job performance in an industrial setting. Educational and Psychological Measurement, 40, 1153-1158.

APPENDIX C

Test Descriptions for Selected Measures Included in the Validity Summary Tables

The references in which these tests appear are listed in Appendix B

SPATIAL ABILITY

Ability to visualize or rotate objects and figures in space.

Space Visualization - ability to visually manipulate or transform the components of a two- or three-dimensional figure to see how things would look under altered conditions.

ASVAB - Space Perception

This measure asks subjects to visualize how cardboard patterns would appear if they were folded along the indicated lines. The test has 20 items with a 12-minute time limit. (The test is no longer part of the operational ASVAB.) (Mathews, 1977)

Factor Referenced Battery - Pattern Comprehension

This is a measure of surface development. Each item consists of a layout pattern, outlined in solid lines and showing folds by dotted lines, together with an isometric drawing of the object that would be made by folding the pattern correctly. The task is to match the dotted lines with the edges of the drawing; the test contains 15 items with a 4-minute time limit. (Curtis, 1971)

Spatial Movement

Each item in the test presents a stimulus pattern or design and four alternative response patterns. The task is to indicate which of the four patterns is the same as the stimulus pattern despite the complications that the matching alternative may be in a different position or folded in some way. (Johnson, Burke, Loeffler & Drucker, 1955)

General Aptitude Test Battery - Three Dimensional Space

Each item contains a three-dimensional figure flattened into two dimensions. The task is to choose, from among several drawings, the one which shows how the figure would look in three dimensions. The test contains 40 items with a 6-minute time limit. (Department of Labor, 1970)

Designs

This test requires the subject to select from a number of parts those parts that will fit together to form the "target" design correctly. Pieces used for the construction may vary from 2 to a maximum of 10. The test contains 22 items and has a time limit of 20 minutes. (Mathews & Jensen, 1977)

Flanagan Industrial Tests - Assembly

This test assesses the ability to visualize the appearance of an object assembled from a number of separate parts. It contains 20 items with a 10-minute time limit. (Flanagan, 1965)

Factor Referenced Battery - Space Perception

This is a test of the ability to mentally invert, rotate, or otherwise manipulate complex stimulus patterns according to explicit directions. There are five block-counting items, five two- dimensional figure-rotation items, four paper-folding and cutting or punching items, and one figure analogy item with a six-minute time limit for all items. (Curtis, 1968)

Two-Dimensional Mental Rotation - ability to identify a two-dimensional figure when seen at different angular orientations.

Visual Recognition

The task in this test is to match a geometrical design given on the left side of the page with one of five designs given on the right. The test contains 40 items. (Eaton, Bessemer & Kristiansen, 1979)

Primary Mental Abilities - Spatial Relations

This test measures the ability to visualize how objects will appear when rotated in space. The test contains 30 items with a 7-minute time limit. (Science Research Associates, 1965)

Figures

This 20-item, 5-minute test requires the examinee to match the "problem" figure with each of six figures that are either exact reproductions or mirror images of the problem figure. (Martinek, Sadacca & Burke, 1965)

Three Dimensional Mental Rotation - ability to identify a three-dimensional object projected on a two-dimensional plane when seen at different angular orientations either within the picture plane or about the axis in depth.

Rotated Blocks

This test requires the subject to select, from among five choices, the one block that is identical to the "target" block. Each of the five response alternatives is presented from a different angle or side than the "target" block. The test contains 20 items with a 20-minute time limit. (Mathews & Jensen, 1977)

Block Counting

This is a measure that requires the subject to examine to "see into" a three-dimensional pile of blocks and to determine how many pieces are touched by a certain numbered block. The test is divided into two sections with 45 items each, with a time limit of 4 minutes per section. (Mathews & Jensen, 1977)

Employee Aptitude Survey - Space Visualization (Test 5)

This measure contains 50 multiple-choice items with 10 alternatives each. The task is to count the number of blocks touching a designated block within a 5-minute time limit. (Ruch & Ruch, 1980)

Guilford-Zimmerman Aptitude Survey - Visualization

This test measures the ability to manipulate ideas visually. The task is to visualize the movements of an object in space. It contains 40 items with a 10-minute time limit. (Guilford & Zimmerman, 1956)

Spatial Scanning - ability to visually survey a complex field to find a particular configuration representing a pathway through a field.

Electrical Mazes

This is a test of the subject's ability to choose a correct path from among five choices. For each item there is a diagram which consists of a large circle at the top of the picture and five lettered boxes at the bottom. In each box there is a dot marked "S" and a dot marked "F." Lines lead from these points to the other boxes and to the circle, with dots indicating connections between lines. The subject must choose the box which has a connection from the "S" through the circle and back to the "F" in the same box. There are 16 such items. (Hunter & Thompson, 1978)

Spatial Orientation - ability to determine one's bearings with respect to points of a compass and the ability to maintain or establish location relative to landmarks in the environment.

Locations Test

This 48 item visual test consists of four small photographs; each set is accompanied by a large photograph with five lettered locations marked on it. The task is to identify the lettered location in the larger photograph from which each of the four small photographs was taken. of the 12 sets of 4 small photographs are darkened to give a "night" effect. (Eaton, 1978)

Aircraft Orientation

In this test, the examinee is presented with a cockpit view of the ground and must visualize what altitude and position the plane must be in to present such a cockpit view -- climb and bank, and so on. The test contains 28 items with a 12-minute time limit. (Martinek et al., 1965)

PERCEPTUAL SPEED AND ACCURACY

Ability to perceive visual information quickly and accurately and to perform simple processing tasks with it (e.g., comparisons).

ASVAB-Coding Speed

Ability to quickly and accurately assign coded numbers is tested by relating them to specific words. The test contains 100 items with a 7-minute time limit. (Campbell & Black, 1982; Jensen & Valentine, 1976)

ASVAB-Attention to Detail

This test is similar to ASVAB Coding Speed. Subjects count the number of letter "c"s in a row of letter "o"s. There are 60 items with a 4-minute time limit. (Greenstein & Hughes, 1977)

Factor-Referenced Battery - Perceptual Speed

Consists of 60 rows of 30 digits each. The left digit in each row is circled. The task is to count all digits in the row that are the same as the circled digit. There are 60 items with a 3-minute time limit. (Curtis, 1968)

Lateral Perception

Ability to discriminate the similarities or differences between letter, number, or symbol patterns is tested. Items consist of two rows of one to ten alphanumeric characters or keyboard symbols. Rows are presented side by side with differing degrees of left-right separation between rows. Subjects must compare the two rows and respond whether the rows are the same or different. (Eaton et al., 1979)

Factor-Referenced Battery - Answer Sheet Marking

This is a test of how quickly and accurately the subject can mark answers. The items are pairs of numbers, and each pair stands for one space on the answer sheet. The first number is the number of the question and the second is the number of the space to blacken for that question. There are two separately timed sections in the test, each containing 75 items, and a total 2-minute time limit. (Hunter & Thompson, 1978)

Basic Test Battery - Clerical

The test consists of 210 number matching items with a 10-minute time limit. Items must be paired according to rules, quickly and accurately. (Cory, 1976)

Counting Numbers

This measures ability to scan rows of digits to identify specified numbers and count their frequencies. (Cory, 1976)

Clerical Carefulness

Subjects are presented with two pages of 49 rows and 15 columns of three-digit numbers. The task is to find the largest number in each row. There are 49 items and a 12-minute time limit. (Osburn, Sheer, Elliott, & Mullins, 1964)

Letter Counting

This test presents the subject with rows of 66 letters arranged randomly. The task is to count the number of letter "g"s in each row. There are 60 items with a 20-minute time limit. (Osburn et al., 1964)

Score Checking

This test requires comparison of printed numbers for their similarity. The test consists of a set of numbers printed on one side of the sheet and a comparison set on the reverse side. There are 400 items and a 28-minute time limit. (Osburn et al., 1964)

Dial Reading

In this test, subjects must read a dial quickly and accurately. There are 30 items and a 4-minute time limit. (Wilbourn & Guinn, 1973)

Paired Letters

The task here is to find, for each item, a pair of letters or figures identical to an underlined pair. The test contains 34 items and has a 3-minute time limit. (Wilbourn, Guinn, & Leisey, 1976)

Number Reversal

In this test, subjects must find the exact reversal of a series of four to seven digits. There are 48 items and a 7-minute time limit. (Wilbourn & Guinn, 1973)

Visual Recognition

This is a 40 item timed test in which the examinee is required to match a geometric design given on the left with one of five geometric designs given on the right. (Eaton, 1978)

General Aptitude Test Battery - Name Comparison (GATB)

Subjects compare two names that may or may not differ slightly, and then judge them to be identical or different. There is a 6-minute time limit for the 150 items on this test. (Department of Labor, 1970)

Speed of Perception

The examinee must locate in succession the numbers from 1 to 50, where the numbers vary in size, location on the page, and orientation. The numbers are presented in random locations on one side of a standard 8 1/2 by 11 inch sheet of paper. (Greenstein & Hughes, 1977)

Flanagan Industrial Tests - Tables

This test measures the ability to read tables quickly and accurately. (Flanagan, 1965)

Flanagan Industrial Tests - Scales

This test measures the ability to read scales, graphs, and charts quickly and accurately. (Flanagan, 1965)

Flanagan Aptitude Classification Test Battery (FACT)

The items in this test are rows of machinery parts, and the subject is required to identify flawed parts. The test contains two sections of 20 items each with a 3-minute time limit for each section. (Osburn et al., 1964)

Factor-Referenced Battery -Table Reading

This is a test of the subject's ability to read tables quickly and accurately. The items consist of pairs of numbers appearing on the abscissa and ordinate of a large table. The subject's task is to find the entry in the table at the intersection of the row and column designated by the pair of numbers. There are five practice problems and 43 scored items in this test. (Osburn et al., 1964)

Factor-Referenced Battery - Scale Reading

This is a test of the subject's ability to read scales, dials, and meters. There are a variety of scales with various points indicated on them by numbered arrows. The subject is to estimate the numerical value indicated by each arrow. There are 24 scored items, divided into two separately timed sections. (Hunter & Thompson, 1978)

Flanagan Industrial Tests - Inspection

This is a measure of ability to spot flaws or imperfections in a series of articles quickly and accurately. (Flanagan, 1965)

Army Perceptual Speed Test

This test requires the subject to match four groups of sketched objects with the proper four or five sketch groups from which they are taken. There are 48 items with a 5-minute time limit. (Greenstein & Hughes, 1977)

Marking Test

Sixteen 10-digit "phone" numbers are presented directly above a representation of a mark sense card, and the task is to mark the numbered boxes that correspond to the 10-digit number presented above the "card." (Gael, Grant, & Ritchie, 1975)

Coding Test

One hundred sets of three letters are presented on a page, and the task is to associate one of three symbols with each set, depending on whether the three letters are the same, whether two are the same, or whether all are different. (Gael et al., 1975)

Perceptual Speed

This test is a 40 x 25 matrix of randomly arranged single digits, in which pairs of like numbers appearing together in a row are to be circled. (Gael et al., 1975)

Short Employment Tests - Clerical Aptitude

This test requires the applicant to locate and verify a name in an alphabetical list, and to read and classify the dollar amount entered opposite that name. Since the task is simple, speed and accuracy are what count. (Bennett & Gelink, 1972)

Employee Aptitude Survey Test 4 - Visual Speed and Accuracy

The subject is required to compare pairs of numbers (and some symbols) and to indicate for each comparison pair whether they are the same or different. There are 150 items to be completed within the 5-minute time limit of this highly-speeded test. (Ruch & Ruch, 1980)

Guilford-Zimmerman Perceptual Speed and Accuracy

Subjects are presented with four stimuli on the left side of the page and five lettered stimuli on the right side. Each stimulus on the left is to be matched with one on the right. There is a 5-minute time limit. (Ronan, 1964)

Primary Mental Abilities (PMA) - Perceptual Speed

This tests the ability to recognize likenesses and differences between objects or symbols quickly and accurately. There are two sections to the test with two different types of items. Each section has 14 items; the first section has a 1 1/2 time limit and the second has a 2-minute limit. (Science Research Associates, PMA Manual, 1962)

Number Size

The task in this test is to determine whether a series of individual numbers is higher or lower than a specified test number. The test contain two parts with 16 items and a 2-minute time limit per part. (Wilbourn & Guinn, 1973)

VERBAL ABILITY

Ability to understand the English Language

Verbal Comprehension - knowledge of the meaning of words.

ASVAB Word Knowledge

This test measures verbal comprehension which entails the ability to understand written and spoken language. The task is to read a statement and then identify the meaning of the underlined word in the text from among four response alternatives. The test contains 35 items with an 11-minute time limit. (Mathews, 1977)

Factor Referenced Battery - Word Knowledge

In this vocabulary test in which the subject chooses the correct synonym for a given word from among four alternatives. The test contains 20 items and a 3-minute time limit. (Curtis, 1968)

Word Knowledge

In this test of how well the subject understands words. Each of the 10 items consists of an underlined word followed by five choices. The subject is to decide which one of the five choices most nearly matches the meaning of the underlined word. (Valentine, 1977)

General Aptitude Test Battery - Vocabulary

Four words are given; the task is to identify two of the four words that represent synonyms or antonyms. The test contains 60 items with a 6-minute time limit. (Department of Labor, 1970)

Vocabulary

In this test the subject is asked to read the first word and then identify from among three alternatives the one that is incorrect or does not mean the same thing as the first word. The test contains 20 items. (Osburn et al., 1964)

Basic Test Battery - General Classification

A 100-item test of opposite, verbal analogy and sentence completion items with a 35-minute time limit. (Thomas & Thomas, 1965)

Flanagan Industrial Tests - Vocabulary

This test measures the ability to choose the right word to convey an idea and knowledge of words used in business and government matters. The test contains 72 items with a 15-minute time limit. (Flanagan, 1965)

Short Employment Tests - Verbal

This 5-minute, 50 item test asks subjects to read each word and then identify from among four alternatives the one word that means the same or most nearly the same. (Bennet & Gelink, 1972)

Personnel Tests for Industry - Verbal Test

Each item contains a question with four response alternatives. These questions ask subjects to identify the word that does not belong, the word that best defines a given word, and so on. The measure contains 50 items with a 5-minute time limit. (Wesman & Doppelt, 1969)

Employee Aptitude Survey - Verbal Comprehension (Test 1)

The task is to read each word and then identify the one word from among four that is the same or about the same as the target word. The test contains 30 items with a 5-minute time limit. (Ruch & Ruch, 1980)

Air Force Reading Abilities Test (AFRAT) - Vocabulary

This measure consists of 45 vocabulary items which asks subjects to identify the correct synonym from among several alternatives. (Mathews & Roach, 1983)

Reading Comprehension - the ability to read and understand written material.

ASVAB - Paragraph Comprehension

Subjects are asked to read a paragraph and then answer a question about the material read. The measure contains 15 paragraphs and questions with a 13-minute time limit. (Campbell & Black, 1982)

Army Classification Battery - Reading and Vocabulary

The 56 item, 25-minute test requires the examinee to read several paragraphs and answer questions pertaining to the meaning of the paragraph and of certain words used in the paragraph. (Helme & White, 1958)

Technical Manual Use Test

In this test, subjects are given a technical manual and are asked to locate information in the index and on given pages and in given sections. The test contains a total of 13 items. (Campbell & Black, 1982)

Science Research Associates - Reading Index

This test of general reading achievement is designed to measure the ability to recognize and decode words and to comprehend phrases, sentences, and paragraphs. (Science Research Associates, 1974)

Primary Mental Abilities - Verbal Meaning

This test measures the ability to understand ideas expressed in words. The test contains 60 items with a 4-minute time limit. (Science Research Associates, 1962)

REASONING

Ability to discover a rule or principle and apply it in solving a problem.

Inductive Reasoning - ability to form and apply hypotheses that fit a set of data.

Factor Referenced Battery - Induction

This test measures the ability to find the general concepts that fit particular sets of data. Subjects are presented with four groups of letters. The task is to discover the rule that relates three of the groups but not the fourth. The test contains 30 items with a 3-minute time limit. (Curtis, 1968)

Employee Aptitude Survey - Numerical Reasoning (Test 6)

Subjects are presented with a series of numbers. The task is to discover the pattern and then to identify the next number in the series. The test contains 20 items with a 5-minute time limit. (Ruch & Ruch, 1980)

Factor Referenced Battery - Letter Sets

This test consists of five groups of letters, each with four letters in each group. Four of the groups of letters are alike in some way. The subject is to find the rule that makes the four groups alike and then identify the one group that does not fit the rule or that is different. The test contains 30 items. (Valentine, 1977)

Deductive Reasoning - ability to use logic and judgment in drawing conclusions from available information.

Flanagan Industrial Tests - Judgment and Comprehension

This test measures the ability to read with understanding, to reason logically, and to use good judgment in interpreting materials. The test has a 15-minute time limit. (Flanagan, 1965)

Factor Referenced Battery - Deduction

This test contains simple syllogisms to assess the ability to reach logical conclusions from given premises. Each item consists of two premises and three alternative conclusions from which the correct conclusion is to be chosen. It contains 15 items with a 2 1/2-minute time limit. (Curtis, 1971)

Nonsense Syllogisms

Subjects are presented with formal syllogisms involving nonsensical content to avoid reference to past learning. Some of the stated conclusions follow correctly from the premises and some do not. The task is to indicate whether or not the conclusion is logically correct. The test contains two parts, each with 15 items and a 4-minute time limit. (Cory, 1976)

Inference

The task for this test is to select one of five conclusions that can be drawn from each previous statement. The test contains two parts, each with 10 items and a 6-minute time limit. (Cory, 1976)

Analysis Aptitude Test

This test measures reasoning and analytical skills in a multiple choice format. Subjects are presented with information and must draw conclusions from it. The test contains 22 items with a 45-minute time limit. (Mathews, 1977)

Employee Aptitude Survey - Verbal Reasoning (Test 7)

This test provides subject with a series of factual statements. The task is to read the statements and then determine whether the conclusions drawn about those facts are true, false, or not known. The test contains 30 items with a 5-minute time limit. (Ruch & Ruch, 1980)

Employee Aptitude Survey - Symbolic Reasoning (Test 10)

In this test, subjects are presented with a statement and a conclusion presented in coded or symbol form. After reading each statement, the subject must determine whether the conclusion is definitely true, definitely false, or impossible to determine from the information given. The test contains 30 items with a 5-minute time limit. (Ruch & Ruch, 1980)

Analogical Reasoning - ability to identify the underlying principles governing relationships between parts of words or objects.

Factor Referenced Battery - Verbal Analogies

This is a measure of the ability to determine the relationships between words. In each of 10 items, the subject is provided with one relationship and part of another. The task is to select from among five alternatives the one that best completes a relationship similar to the first one. (Jensen & Valentine, 1976)

Figural Reasoning - ability to generate and apply hypotheses about principles governing relationships among several figures.

Visual Classification

This test contains 50 items with a 15-minute time limit. Each of the items presents a group of five common objects; the task is to select the one object that does not belong with the rest. (Johnson et al., 1955)

Figure Analogies

In this test, the subject is presented with two figures that have a certain relationship to each other. A third figure is presented that has the same relationship to one of five response figures. The task is to discover the relationship between the first two figures and then identify the one figure that has the same relationship to the third figure. The test contains 10 items. (Hunter & Thompson, 1978)

Related Forms

Subjects are presented with two types of model patterns, Type A and Type B, along with three items. For each item, the task is to classify each item or geometric pattern as Type A or Type B. The test contains 28 groups of items for a total of 84 responses required. (Greenstein & Hughes, 1977)

Card Patterns

This test contains playing cards arranged in various patterns or in a particular series. The task is to discover the pattern or series arrangement for each of 50 items within a 20-minute time limit. (Wilbourn & Guinn, 1973)

Dominoes

In this test, dominoes are arranged in numeric patterns or series. The task is to discover the pattern or series in each of 88 items; the time limit is 25 minutes. (Wilbourn & Guinn, 1973)

Pattern Matching

This test contains pictorial problems that require the subject to select the part from among five alternatives that completes a specified pattern. Subjects are asked to complete 38 items within a 20-minute time period. (Mathews & Jensen, 1977)

Abstract Reasoning

In this test, the subject must discover the pattern in a series of figures and then identify the one figure that comes next in the series. (Boone, 1979)

Word Problems - ability to select and organize relevant information to formulate solutions for mathematical problems.

<u>ASVAB - Arithmetic Reasoning</u>

This measure assesses the ability to think through mathematical problems presented in verbal form. It involves discovery and application of general mathematical principles required to arrive at a correct solution to each problem as well as performance of the necessary calculations to attain the solution. The present measure contains 30 items with a 36-minute time limit. (Campbell & Black, 1982)

<u>Basic Test Battery - Arithmetic Reasoning</u>

This measure contains two separately timed parts. The first involves arithmetic computation and includes 20 items with a 12-minute time limit. The second involves arithmetic reasoning and contains 30 items with a 35-minute time limit. (Hoiberg & Pugh, 1978)

NUMERICAL/MATHEMATICAL ABILITY

Ability to solve simple or complex mathematical problems.

Numerical Computation - speed and accuracy in performing simple arithmetic operations such as addition, subtraction, multiplication, and division.

ASVAB - Numerical Operations

This test measures the ability to perform four arithmetic operations-addition, subtraction, multiplication, and division. It contains 50 items with a 3-minute time limit. (Eaton, et al., 1979)

Factor Referenced Battery - Numerical Test

This test measures elementary knowledge of addition, subtraction, multiplication, division, common and decimal, fractions, squares, cubes, and square root. It contains 15 items with a 6-minute time limit. (Curtis, 1968)

Descriptive Test of Mathematics Skills

This includes four tests: arithmetic skills, elementary algebra, intermediate algebra, and functions and graphs. (Suddick & Bower, 1982)

Personnel Tests for Industry - Numerical

Subjects are required to compute the solution for addition, subtraction, multiplication, and division items, to calculate percentages, measurement of length, area and volume and manipulate decimals and fractions. Solutions are recorded on the test form. The test contains 30 items with a 20-minute time limit. (Wesman & Doppelt, 1969)

Flanagan Industrial Tests - Arithmetic

This test was designed to measure the ability to work quickly and accurately with numbers--to add, subtract, multiply, and divide. The test contains 60 items with a 5-minute time limit. (Flanagan, 1965)

Science Research Associates - Arithmetic Index

This is a test of basic computational ability designed to measure the ability to do fundamental operations with whole numbers, fractions, and mixed numbers, and to successfully manipulate decimals and percents. (Science Research Associates, 1974)

Short Employee Tests - Numerical

This is a written test of simple computations involving addition, subtraction, multiplication, and division. The test contains ninety items with a 5-minute time limit. (Wesman & Doppelt, 1969)

Employee Aptitude Survey - Numerical Ability (Test 2)

This test was designed to measure skill in the four fundamental operations of addition, subtraction, multiplication, and division. Integers, decimal fractions, and common fractions are included in separate tests that are separately timed. Part one has a 2-minute time limit; part two, a 4-minute time limit, and, part three a 4-minute time limit. The total test contains 75 items. (Ruch & Ruch, 1980)

General Aptitude Test Battery - Computation

The test asks subjects to perform addition, subtraction, multiplication, and division in 50 multiple-choice items with a six-minute time limit. (Department of Labor, 1970)

General Aptitude Test Battery - Arithmetic Reasoning

This test contains 25 arithmetic word problems. Subjects are asked to solve these problems within a 7-minute time limit. (Scores on the Computation and Arithmetic Reasoning tests are used to form the Numerical Composite for the GATB validity analyses.) (Department of Labor, 1970)

Primary Mental Abilities - Number Facility

This test measures the ability to work with numbers, to handle simple quantitative problems rapidly and accurately, and to understand and recognize quantitative differences. The test contains 30 items with a 10-minute time limit. (Science Research Associates, 1962.)

Use of Formulations and Number Problems

Ability to use algebraic equations to solve number problems.

ASVAB - Mathematical Knowledge

This test measures functional ability in the use of learned mathematical relationships such as knowledge of algebra, geometry, fractions, decimals, and exponents. The test contains 25 items with a 24-minute time limit. (Mackie, Ridihalgh, & Schultz, 1981)

MEMORY

Ability to recall previously learned information or concepts.

Associative or Rote Memory - ability to recall one part of a previously learned but unrelated item pair when the other part of the pair is presented.

Area Codes Test

This test contains a table listing several cities within states and their associated area codes. Subjects are then presented with a list of the cities and the area codes presented in random order. The task is to associate the correct area code with the correct city. The test contains 84 items with a 6-minute time limit. (Gael, Grant & Ritchie, 1975)

Factor Referenced Battery - Associate Memory

Subjects are given 3 minutes to memorize items pairs. After this period, they are given one member of each pair and are asked to recall the other member. The test contains 21 items and allows 3 minutes for the recall period. (Curtis, 1968)

Object Number

This measure, adapted from the ETS Kit, asks subjects to examine word-number pairs for 3 minutes. After this period, they are presented with the word and must recall the corresponding number. The test contains two sections each with 15 items; the recall time period is 2 minutes per section. (Cory, 1976)

Flanagan Industrial Tests - Memory

In this test, subjects are given 5 minutes to study a word list that pairs familiar words with unfamiliar ones. Subjects are then given 5 minutes to recognize the familiar word associated with the unfamiliar word. Total time is 10 minutes for the 40-item test. (Flanagan, 1965)

Memory Span - ability to recall a number of distinct elements for immediate reproduction.

Coding

This is a symbolic substitution test involving five figures that correspond to response categories on the answer sheet. The test contains 120 items with a 3-minute time limit. (Wilbourn & Guinn, 1973)

Visual Memory - ability to remember the configuration, location, or orientation of figural material.

Visual Memory

This 20-item test requires the subject to first commit to memory each design in a matrix of 20 different geometrical designs. The matrix is then removed and the subject is asked to view 20 rows each containing designs similar to those viewed in the matrix. In each row the subject must locate the one design that appeared in the matrix. (Greenstein & Hughes, 1977)

Factor Referenced Battery - Pattern Detail

Subjects are given 5 minutes to study five abstract patterns. After this period, subjects are given 15 items in which they must identify the one alternative from among five that was presented on the study page. (Hunter, 1975)

PERCEPTION

Ability to perceive a figure or form which is only partially presented or which is embedded in another form.

Flexibility of Closure (Field Independence) - ability to "hold" a given percept or configuration in mind so as to disembed it from other well- defined or complex material.

Hidden Figures

This is a test of perception, visual distraction (Johnson et al., 1955). Test of the subject's ability to see a simple figure in a complex drawing. At the top of each page are five figures, and below these are some numbered drawings. The subject is to determine which lettered figure is contained in each of the numbered drawings. (Hunter & Thompson, 1978)

Flanagan Industrial Tests - Components

This test measures ability to locate and identify important parts of a whole. This involves an ability to change visual patterns, especially flexibility in shifting from a comprehensive pattern to a detailed part. (Flanagan, 1965)

Educational Testing Services - Hidden Patterns

This is a test of ability to recognize simple patterns in complex patterns. Each item consists of a given geometric pattern in which a single configuration is embedded. The task is to mark, for each pattern, whether or not the configuration occurs. The test contains two parts each with 200 patterns and a 3-minute time limit. (Cory, 1976)

General Aptitude Test Battery - Form Perception

Two measures are used to assess the ability to perceive pertinent details in objects or in pictorial or graphic material and to see slight differences in shapes and shadings of figures and widths and lengths of lines. The Tool Matching Test includes 79 items with two 5-minute time limits and the Form Matching includes 60 items with a 6-minute time limit. (Department of Labor, 1970)

Speed of Closure - ability to identify words or objects given sketchy or partial information.

Educational Testing Services - Gestalt Completion

Perceptual closure in recognition of objects from fragmentary details is measured. Drawings are presented which are composed of black blotches representing parts of the objects portrayed. The subject writes down the name of the object. The test contains two parts with ten pictures and a 2-minute time limit per part. (Cory, 1976)

Concealed Words

Perceptual closure in recognition of words from fragmentary details is measured. Words are presented with parts of each letter missing. The subject is to write out the word in an adjacent space. The test contains two parts with 25 words in each part. Subjects are allowed four minutes per part. (Cory, 1976)

Object Completion

This tests ability to detect a partially obscured outline. Subjects are required to identify a set of partially obscured line drawings or military objects such as field glasses, canteens, etc. (Eaton, 1978)

Hidden Objects

Pictures are presented in which there are hidden or camouflaged objects. The subject is to find the objects within the pictures. (Egbert, Meeland, Cline, Forgy, Spickler & Brown, 1958)

Precision Counting

The task is to count the number of symbols contained in a pictorial item. There are 50 items with a 4-minute time limit. (Wilbourn and Guinn, 1973)

Estimation of Length and Size - ability to use stimuli in the environment to estimate the size or weight of objects or distance between objects.

Point Distance

The subject is required to compare small distances rapidly. Each item has a marked central point surrounded by lines and curves, among which there are dots labeled "a" and "b." The examinee must quickly decide which of the two lettered dots is nearer to the central point. The test is divided into two sections, with 300 items per section and a 2-minute time limit per section. (Hunter & Thompson, 1978)

Simulated Zeroing

A test was constructed to determine the extent to which the subject is able to locate the geometric center of a hypothetical three-round shot group. The score is a mesaure based upon the deviation of perceived center from true center. (Eaton et al., 1979; Greenstein and Hughes, 1977)

Perceptual Discrimination

Subjects must arrange 10 diamonds in descending order of size. There are 21 items with a 25-minute time limit. (Osburn et al., 1964)

Estimation of Length

Subjects are presented with a line and are asked to estimate its length by comparison with a standard set of five lines. There are 120 items with a 12-minute time limit. (Osburn et al., 1964)

FLUENCY

Ability to rapidly generate words or ideas related to target stimuli.

Associational Fluency - ability to rapidly produce words that share a given area of meaning or some other semantic property.

No measures were included in the validity analyses from this area.

Expressional Fluency - ability to rapidly think of word groups or phrases.

No measures were included in the validity analyses from this area.

Ideational Fluency - ability to write a number of ideas about a given topic or examples of a given class of objects.

Factor Referenced Battery - Ideational Fluency

Subjects are asked to think of and list the names of as many things as possible "that are round or that could be called round" within a 3-minute time limit. (Curtis, 1968)

Flanagan Industrial Tests - Ingenuity

This is a test of the ability to think of clever and effective ways of doing things. Subjects are presented with a problem along with clues about how to solve the problem. In addition, five response alternatives hint at the solution by providing the first and last letter in each word of the correct solution. Subjects are given 15 minutes to read and identify a solution for 20 problems. (Flanagan, 1965)

Word Fluency - ability to produce words that fit one or more restrictions that are not relevant to the meaning of words.

Employee Aptitude Survey - Word Fluency (Test 8)

This test is designed to measure the ability to rapidly think of words. Subjects are given a letter such as "S" and are asked to generate as many words as possible in a 5-minute period. (Ruch & Ruch, 1980)

MECHANICAL APTITUDE

Ability to perceive and understand the relationship of physical forces and mechanical elements in a prescribed situation.

ASVAB Mechanical Comprehension

Ability of subjects to determine the operating characteristics of mechanical devices is measured. It requires understanding of mechanical principles underlying the operation of such devices as gears, pulleys, and hydraulic systems. This test has 25 items with a 19-minute time limit. (Jensen & Valentine, 1976)

Factor Referenced Battery - Mechanical Information

This test measures knowledge of tools, tool functions, and mechanical principles. There are 30 items with a 3-minute time limit. (Curtis, 1968)

Basic Test Battery - Mechanical Test

This test contains two separately timed parts: Tool Knowledge has 50 items with a 10-minute time limit. Mechanical Comprehension has 50 items with a 25-minute time limit. (Curtis, 1968)

Mechanical Abilities

This test of knowledge about general mechanics and tool functions contains two parts. Part 1 has statements about general mechanics such as automotives or other mechanical objects. There are 30 items. Part 2 requires the subject to identify uses of tools presented in pictures, and there are 20 items. (Eaton et al., 1979)

Mechanical Principles

Contains 10 items covering mechanical principles and devices, such as gears and pulleys. (Hunter & Thompson, 1978)

Wheels

The task is to determine the direction of a series of wheels when the direction of one wheel in the series is given. There are 60 items with a 10-minute time limit. (Wilbourn & Guinn, 1973)

Flanagan Industrial Test - Mechanics

Ability to understand mechanical principles and to analyze mechanical movements is evaluated. There are 30 multiple-choice items and a 15-minute limit on this test. (Flanagan, 1965)

Differential Aptitude Battery - Mechanical Reasoning

Each item consists of a pictorially presented mechanical situation together with a question about the picture. The examinee is asked to complete 70 items in 30 minutes. (Bennet, Seashore, & Wesman, 1973)

Bennett Mechanical Comprehension Test

This test is very similar to the DAT - Mechanical Reasoning Test. The examinee is presented with a picture depicting a mechanical situation along with a question about the picture. The test contains 68 items with a 30-minute time limit. (Bennet, 1969)

AUTO, SHOP, AND TOOL KNOWLEDGE

Knowledge of automobiles, shop practices and tools, and their uses.

ASVAB-Automotive/Shop Information

This measure assesses knowledge and understanding of automobiles and of tool and shop practices. The test contains 25 multiple-choice questions with an 11-minute time limit. (Campbell & Black, 1982)

ASVAB-Automotive Information

General knowldge about automobiles and automobile engines is assessed. The test contains 25 items. (Eaton et al., 1979)

ASVAB-Shop Information

This test assesses previous knowledge of shop practices and the use of tools in specific situations. There are 25 items. (Jensen & Valentine, 1976)

Basic Test Battery - Shop Practices

This 30-item test covers knowledge of tools and shop equipment. (Cory, 1976)

Factor-Referenced Battery - Tool Functions

Questions about the use of tools are presented. In each of the 10 items, a tool is depicted and five statements are given concerning the use or type of the tool. The subject must select the statement that best fits the illustration. (Hunter & Thompson, 1978)

Factor-Referenced Battery - Tools

This is a test about tools and how they are used. Each of the 10 items has a picture of a tool and four other objects. The subject must decide which one of the four objects goes with the pictured tool. (Hunter & Thompson, 1978)

Tool and Object Nomenclature

In this use and recognition test, typical tools and objects from Navy life are presented and briefly discussed. Then the three 15-item true/false subtests are administered. (Cory, 1982)

Note: The remaining tests involve the assessment of knowledge acquired through formal training.

ELECTRONICS / ELECTRICAL KNOWLEDGE

Knowledge of electrical or electronic systems and operations.

ASVAB Electronics Information

This tests ability to apply previously acquired knowledge in the areas of electricity and electronics toward the solution of problems in practical situations, and assesses knowledge of electricity, radio principles, and electronics. The test has 20 items with a 9-minute time limit. (Mackie et al., 1981)

Factor-Referenced Battery - Electrical Information

The subject's knowledge of electricity and electrical devices is tested. It contains 10 items which cover a variety of electrical principles and applications. (Hunter & Thompson, 1978)

Basic Test Battery - Electronics Technician Selection Test

This test measures achievement and knowledge in areas related to electronic maintenance. The test has five subtests, with a total of 80 items and a 75-minute time limit. (Thomas & Thomas, 1965)

Flanagan Industrial Test - Electronics

This test measures ability to understand electrical and electronic principles and to analyze diagrams of electrical circuits. The test contains 30 items and has a 15-minute time limit. (Flanagan, 1965)

SCIENCE KNOWLEDGE

Knowledge of basic scientific principles.

ASVAB-General Science

This test assesses knowledge of physical and biological sciences. It contains 25 items with an 11-minute time limit. (Campbell & Black, 1982)