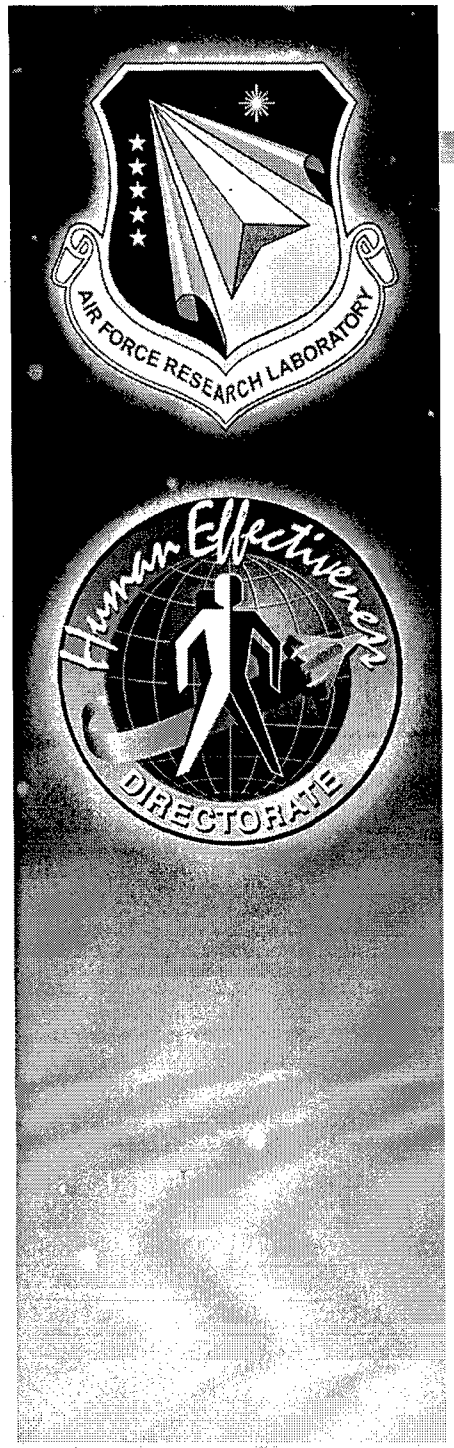
AFRL-HE-WP-TP-2005-0028



AIR FORCE RESEARCH LABORATORY

Tests of Cognitive Ability

Malcolm James Ree
Our Lady of the Lake University
San Antonio TX

Thomas R. Carretta
Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433-7022

December 2005

Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
Dec-2005	Technical Paper	

4. TITLE AND SUBTITLE
Tests of Cognitive Ability

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
*Malcolm James Ree
**Thomas R. Carretta

5d. PROJECT NUMBER
7184

5e. TASK NUMBER
09

5f. WORK UNIT NUMBER
72

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)
*Our Lady of the Lake University
San Antonio TX

8. PERFORMING ORGANIZATION REPORT NUMBER
AFRL-HE-WP-TP-2005-0028

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
**Air Force Materiel Command
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433-7022

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/HECV

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
This is a book chapter. Clearance No. AFRL/WS-05-2719, 1 Dec 05

14. ABSTRACT

This chapter consists of six parts. Part one briefly reviews the historical foundation of the concept of cognitive ability and early attempts to measure it. Part two reviews modern theories of the structure of cognitive ability and the emergence of the concept of general cognitive ability. Next, part three introduces the concepts of specific abilities, knowledge, and noncognitive traits. Part four discusses psychometric characteristics of tests including reliability and validity. Part five reviews the issues to be considered when deciding whether to choose from among commercially-available tests or develop a test. Example questions to help in test construction are provided. The sixth and final part is a general summary.

15. SUBJECT TERMS
Cognitive ability, knowledge

16. SECURITY CLASSIFICATION OF: Unclassified			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Thomas R. Carretts
UNC	UNC	UNC	SAR	63	19b. TELEPHONE NUMBER *(include area code)* (937)255-656-7014

TESTS OF COGNITIVE ABILITY

Malcolm James Ree

*Our Lady of the Lake University, San Antonio, T*exas

Thomas R. Carretta[1]

Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio

OVERVIEW

This chapter consists of six parts. Part one briefly reviews the historical foundation of the concept of cognitive ability and early attempts to measure it. Part two reviews modern theories of the structure of cognitive ability and the emergence of the concept of general cognitive ability. Next, part three introduces the concepts of specific abilities, knowledge, and noncognitive traits. Part four discusses psychometric characteristics of tests including reliability and validity. Part five reviews the issues to be considered when deciding whether to choose from among commercially-available tests or develop a test. Example questions to help in test construction are provided. The sixth and final part is a general summary.

HISTORICAL FOUNDATIONS

The concept of cognitive ability can be traced back over 2,500 years. Zhang (1988) reported that in the sixth century BC, the great Chinese philosopher Confucius, divided people into three groups based on intelligence: people of great wisdom, people of average intelligence, and people of little intelligence. Another Chinese philosopher, Mencius (fourth century BC), likened intellectual measurement to measurement of physical properties. Within a century, the Han dynasty (202 BC -200 AD) had heeded Confucius and Mencius and implemented a system of civil service tests in China.

In the fourth century BC, Aristotle made a distinction between ability (*dianoia*) and emotional and moral capacity (*orexis*). Zhang (1988) reported on the custom of testing children at one year of age beginning in the sixth century AD in China, particularly in southern China. This was described in the writings of Yen (531-590 AD). Zhang (1988) also noted that the use of puzzles to test cognitive ability was popularized during the Song dynasty (960-1127 AD). One example consisted of several geometric shapes that could be manipulated and fit into a variety of designs. The test was designed to measure creativity, divergent thinking, and visual-spatial perception. Another popular Chinese puzzle test designed to measure reasoning ability consisted of interconnected copper rings

mounted on a bar with a rod running through their center. The goal of the test was to remove the bar from the center of the rings.

In the west, the examination of human cognitive abilities was taken up by religious philosophers. In the 16th century AD, Descartes, the French secular philosopher regarded ability as *res cogitans*, the thing that thinks.

In 1575, Juan Huarte published in Spanish (Peiró & Munduate, 1994) a treatise on work and human ability called *Examen de Ingenios*. It was later published in English as *The examination of men's wits: Discovering the great differences of wits among men and what sort of learning suits best with each genius*.

The modern scientific study of human cognitive abilities, however, is often attributed to Binet in France and to the World War I Army Alpha and Beta tests in America.

GENERAL COGNITIVE ABILITY

The English polymath, Sir Francis Galton (1869), invented the construct of general cognitive ability, calling it *g* as shorthand. Charles Spearman (1927, 1930) made the concept of *g* more accessible to psychology through his two-factor theory of human abilities which proposed that every measure of ability had two components, a general component (*g*) and a specific component (*s*).

While the general component was measured by every test, the specific component was unique to each test. Though each test might have a different specific component, Spearman also observed that *s* could be found in common across a limited number of tests. Therefore, the two-factor theory allowed for a spatial factor or other factor that was distinct from *g* but could be found in several tests. These factors shared by tests were called "group factors." Spearman (1927) identified several group factors and noted (Spearman, 1937) that group factors could be either narrow or broad. He further observed that *s* could not be measured without measuring *g*. As we have written elsewhere (Ree & Carretta, 1996, 1998),

> To be accurate, we should call mathematics not M but
>
> $$g + \mathrm{M}$$
>
> with *g* written large to indicate its contribution to the variance of the factor. (Ree & Carretta, 1996, p. 113).

In fact, tests that do not even appear to measure *g* do so as illustrated by Rabbitt, Banerji, and Szymanski (1989) who demonstrated a strong

correlation (.69) between "Space Fortress" a psychomotor task that looks like a video game, and an IQ test.

Controversy about *g* has not abated despite Spearman's early assertion (1930) that *g* was beyond dispute. In contrast to Spearman's model, Thurstone (1938) proposed a multiple ability theory. Thurstone allowed no general factor, only seven "unrelated abilities" that he called "primary." Spearman (1938) reanalyzed Thurstone's data noting that *g* had been submerged through rotation. He then demonstrated the existence of *g* in Thurstone's tests. This finding was independently confirmed by Holzinger and Harmon (1938) and finally by Thurstone and Thurstone (1941). Despite empirical evidence, theories of multiple abilities held sway (Fleishman & Quaintance, 1984; Gardner 1983; Guilford, 1956, 1959; Sternberg, 1985). This was particularly true in psychometrics, where these theories lead to the construction of numerous multiple ability tests such as the Differential Aptitude Test, General Aptitude Test Battery, Armed Services Vocational Aptitude Battery, Air Force Officer Qualifying Test, Flanagan Aptitude Tests, Flanagan Industrial Tests, and others. Cleaving to the empirical data, other researchers continued to study *g* (Arvey, 1986; Gottfredson, 1986, 1997; Gustafsson, 1980, 1984, 1988; Jensen, 1980, 1993, 1998; Schmidt & Hunter, 1998, 2004; Thorndike, 1986; Vernon, 1950, 1969).

Fairness and Similarity: Near Identity of Cognitive Structure

There are several issues that must be addressed when measuring ability in sex and ethnic groups. One of these is that the same factors should be measured for all groups. McArdle (1996) among others has advocated that factorial invariance (i.e., equality of factor loadings) should be demonstrated before other group comparisons (e.g., mean differences) are considered. McArdle stated that if factorial invariance is not observed, the psychometric constructs being measured may be qualitatively different for the groups being compared, obscuring the interpretation of other group comparisons.

Several studies of cognitive factor similarity have been conducted. Comparing the factor structure of World War II U. S. Army pilot selection tests for Blacks and Whites, Michael (1949) found virtually no differences. Humphreys and Taber (1973) also found no differences when they compared factor structures for high and low socio-economic status boys from Project Talent. Although the ethnicity of the participants in Project Talent was not specifically identified, they expected that the ethnic composition of the two groups would differ significantly.

Using 15 cognitive tests, DeFries, Vandenberg, McClearn, Kuse, Wilson, Ashton, and Johnson (1974) compared the structure of ability for Hawaiians of either European or Japanese ancestry. They found the same four factors and nearly identical factor loadings for the two groups.

These studies all examined common factors. Using a hierarchical model, Ree and Carretta (1995) examined the comparative structure of ability across sex and ethnic groups. They observed only small differences on the verbal/math and speed factors. No significant differences were found for *g* on ability measures.

Carretta and Ree (1995) made comparisons of aptitude factor structures in large samples of young Americans. The factor model was hierarchical including *g* and five lower-order factors representing verbal, math, spatial, aircrew knowledge, and perceptual speed. The model showed good fit and little difference for both sexes and all five ethnic groups (White, Black, Hispanic, Asian-American, and Native-American). Correlations between factor loadings for the sex groups and for all pairs of ethnic groups were very high, approaching, $r = 1.0$. Comparisons of regression equations between pairs of groups indicated that there was no mean difference in loadings between males and females or among the ethnic groups. These and previous findings present a consistent picture of near identity of cognitive structure for sex and ethnic groups.

Predictive Fairness

Several researchers have conducted studies of predictive fairness of cognitive ability tests. Jensen (1980) noted that numerous large scale studies provided no evidence for predictive unfairness. He concluded that predictive bias

did not exist, although intercept differences could be observed and were likely due to sampling error or differences in reliability for the two groups (p. 514).

Putting a finer point on it, Carretta (1997) demonstrated that even when intercept differences were observed in statistical tests of differences of regression equations for two groups, the differences were due solely to differing reliability found in the two groups

Hunter and Schmidt (1979) investigated 39 studies of Black-White validity and found no evidence of differential prediction for the groups. Schmidt and Hunter (1982) illuminated pitfalls in assessing the fairness of regressions using tests of differences in regression (linear) models. In these two studies, Hunter and Schmidt concluded that artifacts accounted for the apparent differential prediction and that no predictive bias was present. Carretta (1997) and Jensen (1980) provide clear statistical explanations of the issues.

In sum, no evidence exists that cognitive ability tests are unfair.

SPECIFIC ABILITY, KNOWLEDGE, AND NONCOGNITIVE TRAITS

The measurement of specific abilities, knowledge, and noncognitive traits often has been proposed as crucial for understanding human characteristics and occupational performance. Ree and Earles (1991) have demonstrated the lack of predictiveness for specific abilities while Ree and others (Olea & Ree, 1994; Ree,

Carretta, & Doub, 1998/1999; Ree, Carretta, & Teachout; 1995; Ree, Earles, & Teachout, 1994) demonstrated the predictiveness of job knowledge.

McClelland (1993), for example, suggested that under some circumstances noncognitive traits such as motivation may be better predictors of job performance than cognitive abilities. Sternberg and Wagner (1993) proposed the use of measures of tacit knowledge and practical intelligence in lieu of measures of "academic intelligence." They define tacit knowledge as "the practical know how one needs for success on the job" (p. 2). Practical intelligence is defined as a more general form of tacit knowledge. Schmidt and Hunter (1993), in a review of Sternberg and Wagner, note that their concepts of tacit knowledge and practical intelligence are redundant with the well established construct of job knowledge. Additionally, Ree and Earles (1993) pointed out the lack of rigorous empirical evidence to uphold the assertions of McClelland, Sternberg, and Wagner as well as other critics.

The construct of Emotional Intelligence (Goleman, 1995) has been proposed as another facet that is more important than ordinary cognitive ability. Although its proponents (e.g., Mayer, Salovey, & Caruso, 2002) consider it to be a distinct construct, Schulte, Ree, and Carretta (2004) have demonstrated that it is not much more than a combination of the existing constructs of cognitive ability and personality.

PSYCHOLOMETRIC CHARACTERISTICS OF MEASURES OF COGNITIVE ABILITY

Courses in statistics and research methods are common for human resources/personnel specialists and there are established guidelines for conducting studies of personnel measurement and selection (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1999; Society for Industrial-Organizational Psychology, 2003). Reliability and validity are two core concepts that must be considered whether choosing a commercial test or developing a test.

Reliability

Reliability is best defined as precision of measurement; that is how much of the measurement is true and how much is error. In this statistical context "error" does not mean wrong, but random fluctuation. An error has not been committed, rather random fluctuation happens *per force* and cannot be avoided although it can be minimized. From this basic definition flow the other popular definitions of reliability such as stability over time and consistency across test forms, as well as internal consistency. Stability over time typically is measured by retesting people after a period of time to ensure that their scores are consistent (i.e., test-retest reliability). Stability across test forms measuring the same

construct(s) is referred to as alternate form reliability. Internal consistency is measured by assessing the extent to which items are correlated with each other (e.g., correlating odd items with even items or split-half reliability or coefficient alpha). All three of these indices of reliability are typically measured using correlations or approximations to correlations. Although correlations usually range from +1.0 to -1.0, a reliability coefficient is a ratio of true variance to total variance ..... .

Two widely used cognitive ability tests are the Wonderlic Personnel Test and the Watson-Glaser Critical Thinking Appraisal. According to research cited in the *Wonderlic Personnel Test & Scholastic Level Exam User's Manual*, the test-retest reliability ranges from .82 to .94; alternate form reliability ranges from .73 to .95; and split-half reliabilities range from .88 to .94. Similarly high levels of reliability are noted in the *Watson-Glaser Critical Thinking Appraisal Manual (Form S)*. Test-retest reliability was .81 for a sample of 42 employees and internal consistency reliabilities ranged from .66 to .87 in a wide variety of jobs. The data from these two well-known and frequently used tests shows that cognitive ability is a reliably measured construct.

For a test to be reliable there must also be consistent administration, consistent collection of answers, and objective scoring. Test administration procedures must not vary from examinee to examinee and the data collection

methods must be consistent. For example, Ree and Wegner (1990) showed that apparently minor changes in machine-scored answer sheets could produce major changes in tests scores, particularly in speeded tests. This issue looms larger as we consider placing our test for the selection of applicants on a computer where the presentation could vary by screen size, contrast, and font type. Additionally, when different administration modes or response collection are necessary, it is essential to develop statistical corrections for the scores (Carretta & Ree, 1993). The use of tests of poor reliability to make decisions about excluding applicants, especially applicants near the minimum cutting point from a training program, is bad practice and may lead to indefensible consequences in cour,t should a legal challenge arise.

Scoring must be objective. A correct answer must be counted correct by all scorers. To deviate from this will cause scores to vary by who did the scoring and will reduce reliability of the test leading to reduced validity and possibly an indefensible position in court. This is less of a problem for a multiple choice test where the answer is presented and must be identified from among answers presented. It is more of a problem for an essay type exam where the answer must be produced and evaluated.

General cognitive ability can be reliably measured through several methods. Because it is the greatest source of variance in cognitive tests it is

relatively easy to get acceptable reliability by careful item construction and by adding items. However, as Thompson (2003) has pointed out, the reliability to be considered is the reliability in the sample currently being investigated, not that from previous test administrations or the normative sample.

> "It is important to evaluate score reliability in *all* (emphasis in original) studies, because it is the reliability of the data in hand that will drive study results, and not the reliability of the scores described in the test manual..." (Thompson, 2003, p. 5).

Validity

The important question about validity is whether a test measures what it claims to measure. Although it is convenient to distinguish several types of validity, the argument can be made that all validity studies are really construct validity studies. If the test can be shown to be valid it is shown to be measuring the construct and therefore construct validity is bolstered.

However, a caveat must be offered here. A measure can have predictive validity where it is assumed that it measures a certain construct, but in fact measures a different construct. For example, Walters, Miller, and Ree (1993) in a validation of a structured pilot candidate selection interview, reported validity for

training performance. At first glance, it appeared that the validity of the interview came from measuring motivation and job knowledge. However, its lack of incremental validity over a cognitive ability measure suggested that the validity of the interview was, at least to some extent, due to its measurement of cognitive ability. The same may be cited for the example of the psychomotor test "Space Fortress" (Rabbitt, Banerji, & Szymanski, 1989) that looked different than a cognitive ability test, but on analysis was found to be a cognitive ability measure. Elsewhere (Walters et al., 1993), we have identified this as the "topological fallacy."

Schmidt and Hunter (1998) reported on the validity of cognitive ability from a very large meta-analytic study conducted for the U.S. Department of Labor (Hunter, 1980; Hunter & Hunter, 1984). The database for the meta-analysis included over 32,000 employees in 515 widely diverse civilian jobs. Similar to reliability, validity can be assessed using correlations which can range from +1.0 to -1.0. They found that the validity of cognitive ability for predicting job performance was .58 for professional-managerial jobs, .56 for high level complex technical jobs, .51 for medium complexity jobs (which comprises 62% of all jobs in the U.S. economy), .40 for semi-skilled jobs, and .23 for completely unskilled jobs. In sum, the research evidence for the validity of cognitive ability measures for predicting job performance is very strong.

HOW TO SELECT OR DEVELOP A COGNITIVE ABILITIES TEST

Selection or development of a test begins with job analysis. The goal of job analysis (Cascio, 1991; Gael, 1988; McCormick, 1976, 1979) is to establish job, task, and cognitive requirements or Knowledge, Skills, Abilities, and Other (KSAO) requirements. A job analysis should be conducted whether one ultimately chooses to select a cognitive ability test from among commercially available tests or to develop a test. It can be accomplished many different ways. Cascio (1991) provides a discussion of several methods. The results of the job analyses should lead to the development of a structured taxonomy and test specification requirements (e.g., test content, reading level, item difficulty, item discriminability).

Once a job analysis has been completed and a particular set of abilities has been identified as necessary for successful job performance, the next step is to decide whether to select an existing test or develop one to measure those abilities.

Whether the decision is to select an existing test or to develop one, information should be obtained regarding several factors in order to make a reasoned evaluation. These include: 1) test development documentation such as theoretical basis, normative sample, and test development procedures, 2)

psychometric characteristics, such as reliability, validity, and test bias, 3) information regarding administration including materials, procedures, instructions, reasonable accommodation for applicants with disabilities, and special training required, 4) test interpretation aids such as normative data, expectancy charts, or cut scores, 5) scoring options; whether the test is hand scored, computerized, or machine scored and the qualifications for scoring, 6) on-going research/refinement of the test, 7) time requirements, 8) credentials and expertise of test developers, 9) total costs including materials, fees, and test development costs.

The following sections discuss factors that affect the decision to select an existing test or develop one for use in personnel measurement and selection and the associated activities. The entry-level electrician job is used throughout as an example.

Selecting an Existing Cognitive Ability Test

Reasons for Selecting an Existing Cognitive Ability Test

The use of commercially-available tests is often an attractive choice. Two common reasons for using commercially-available tests are: 1) to avoid the costs associated with test development and maintenance and 2) the availability of normative and psychometric data. The level of effort, technical expertise, and

other resources required for test development and maintenance may be prohibitive for many test users.

Test development activities include development of test specifications and item pools, technical and sensitivity reviews, creation of test administration procedures/instructions, trying out items and analyzing item-level data, assembling and producing the test, and preparing test documentation such as test manuals. Test development costs are further exacerbated as it is often desirable to periodically develop new test forms (i.e., combat test compromise, update test content). The availability of normative (population and subgroup performance) and psychometric (e.g., reliability, content validity, construct validity, criterion-related validity) data for commercially-available tests provides valuable information and a context for interpretation of test results (e.g., comparison to other groups).

Identifying Candidate Cognitive Ability Tests

The first step in selecting a commercially-available cognitive ability test is to identify candidate tests. There are several helpful sources for information about commercially-available tests and much information is now available via the internet (e.g., Buros Institute of Mental Measurements, http://www.unl.edu/buros, Educational Testing Service, http://www.ets.org, and Pro-Ed,

http://www.proedinc.com). Example publications include the *Mental Measurement Yearbook* (Plake, Impara, & Spies, 1993), *Tests in Print* (Murphy, Plake, Impara, & Spies, 2002), and *Test Critiques* (Keyser & Sweetland, 1997). The *Mental Measurement Yearbooks,* a set of volumes covering many decades and shelf space include descriptive information, professional reviews, and references. Each volume only includes information about tests that are new, revised, or in wide use since the previous edition. *Tests in Print* is a comprehensive bibliography of all known commercially-available tests in print in the English language. It provides information about the purpose of the test, what it measures, author, publisher, publication date, in-print status, cost, intended test population, and administration time. *Test Critiques* includes tests used in business, education, and psychology. It provides information regarding practical applications and uses, guidelines for administration, scoring, and interpretation, psychometric data (norms, reliability, validity), and critical reviews. Professional organizations such as the American Psychological Association (APA) and the Educational Resources Information Center (ERIC) are other valuable sources of information about commercially-available tests. The APA neither endorses nor sells tests, but provides guidance about locating information about tests and their proper use (http://www.apa.org/science/faq-findtests.html). The same is generally true for the American Psychological Society

(http://www.psychologicalscience.org/) and the Society for Industrial and Organizational Psychology (http://www.siop.org/Workplace/default.htm). The Association of Test Publishers (http://www.testpublishers.org/), a non-profit organization representing providers of assessment tools, tests, and related services is yet another source of information.

Evaluating the Information about Candidate Tests

Once a set of candidate tests has been identified, the next step in selecting a test is to evaluate the available information. A variety of issues should be considered when tests are to be used in an employment context. These include the appropriateness of the test for its intended use, administrative procedures (individual vs. group, paper-and-pencil vs. computerized, administration time, need for special equipment or setting, training of administrators), interpretability of test scores (norms), cost, usefulness of supporting materials (administrative and technical manuals), and psychometric properties (error of measurement, reliability, validity, lack of bias). Fortunately, this information is available from the sources described above. Naturally, while the reviews of published tests are useful, the final evaluation of a particular test's usefulness must be postponed until the test has been obtained, the materials studied, and the test has been administered and validated. Due to the technical nature of the material in test

reviews and manuals, it may be necessary to obtain professional assistance in interpreting test information.

Consider an example regarding the evaluation of applicants for an entry-level electrician job training program. A review of the job analysis results, as required by law, will guide us regarding both the specific abilities required (e.g., verbal comprehension, mathematics, spatial reasoning) and their level (e.g., 10^{th} grade or higher for verbal comprehension and 12^{th} grade or higher for mathematics and spatial). A review of the information provided in the *Mental Measurement Yearbooks*, *Tests in Print*, and *Test Critiques* will allow the user to identify candidate tests that already exist and whether they are appropriate for use in the current context.

Looking in the references cited above, several tests were found that met the content requirements. Among these tests were the *Armed Services Vocational Aptitude Battery, Differential Aptitude Tests,* and the *Wonderlic Personnel Test.* All are professionally developed and have acceptable psychometric properties (i.e., reliability, validity, norms). The *Armed Services Vocational Aptitude Battery* is not available for commercial purchase. Both the *Differential Aptitude Tests* and the *Wonderlic Personnel Test* are available commercially and can be group administered. For purposes of the entry-level electrician job, the *Wonderlic Personnel Test* offers the advantage of lower costs and shorter administration

time. If the job analysis had revealed that content measured by the *Differential Aptitude Tests* and not the *Wonderlic Personnel Test* was required then the *Differential Aptitude Tests* or other similar tests would have been a proper choice.

Obtaining the Test

Once an appropriate test has been identified, the next step is to obtain copies of the test, answer sheets if necessary, test manuals and permission to reproduce or use the test. Requirements for the purchase and use of tests vary across test publishers (Eyde et al., 1993). Some test publishers will permit potential test users to purchase only the test manual, in order to further evaluate the suitability of the test prior to making a final purchase decision. To qualify for test purchase, some test publishers require the purchaser to have an advanced degree in psychology, education, or a related field, complete specialized training in test administration, methodology, and use, and possess a professional license.

Trying Out the Test

After the test has been obtained it must be tried out with the intended applicant target group, including an assessment of its reliability and validity. Reliability can be estimated via test-retest, parallel forms, or internal consistency, if appropriate assumptions can be met (Cronbach, 1951), in the applicant group.

The general standards for validity studies are described in §1607.5 of the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978). During this stage, predictor and criterion measures are identified, data are collected on an appropriate sample, and predictive validity is examined. In a predictive validity design, the appropriate sample is a large group of applicants (i.e., several hundred). The selection instruments are administered during application and the criteria are collected after those selected have completed training or been on the job for some period. A correlation may be computed from the data collected on the predictor test and the criterion. As this correlation is likely downwardly biased due to pre-selection, the correction for range restriction (Hunter & Schmidt, 2004; Ree, Carretta, Earles, & Albert, 1994) should be applied to determine how the validity would be in the full applicant sample.

In addition to the selection test's usefulness for identifying those likely to be successful, there are other important considerations in evaluating the test for personnel selection. For example, it is important to determine whether the way the test is used differentially qualifies members of different subgroups (i.e., adverse impact). Cutting or qualification scores or combining the weighted test scores with other selection information defines the selection decision and the selection decision defines whether groups qualify at differing rates.

Another important factor to consider is whether the test predicts training and job performance equally well for members of different sex and ethnic/racial groups (i.e. predictive bias). Information about studies done to examine adverse impact and predictive bias might be available in the test documentation.

Developing a Cognitive Ability Test

Reasons for Developing a Cognitive Ability Test

Despite the availability of off-the-shelf commercial cognitive ability tests with acceptable psychometric properties, there are several reasons why it may be desirable for organizations to develop new tests. Some reasons are: 1) a proprietary test is desired, 2) alternate forms are required but not available in the commercial test, 3) test content becomes outdated, and 4) there is a need to measure a newly hypothesized or highly specialized ability.

Proprietary tests are desirable when organizations want to control test content, administration and scoring procedures, and testing policy (e.g., test-retest). A well-known example is the *Armed Services Vocational Aptitude Battery* (ASVAB), which is used for US military enlistment qualification. Despite careful efforts to control test exposure, commercially available ASVAB study guides are readily available and there is a constant potential threat to compromise of test content.

Alternative forms include parallel forms (different items, but equivalent content and score distributions) and forms administered in different formats (paper-and-pencil, computer-administered, computerized adaptive test). Alternative forms are useful when retests are allowed and to combat possible test compromise. In the case of retests, when alternate forms are available applicants can be retested on a form with different, but construct and psychometrically equivalent, items to reduce retest gains due to prior exposure to test items.

Test content may become outdated for several reasons. Two examples are when word usage patterns change or when test content regards technological areas that change rapidly. New words enter common usage while others drop out over time. An example of outdated test content might be items from a 1960's vintage electrical knowledge test that included questions about vacuum tubes and audio output transformers. These items clearly would not be appropriate for measuring knowledge about state-of-the-art electrical technology.

Sometimes it is desirable to develop test content based on newly hypothesized ability taxonomies if they were found in the job analysis (e.g., procedural knowledge, working memory capacity) or specialized content not found in commercial off-the-shelf cognitive tests. Specialized technical content tests are common in the US military. For example, the ASVAB includes subtests that measure knowledge of electricity and electronics, mechanical and physical

principles, automobile terminology and technology, and tools and shop terminology and practices. Similar commercial tests are available and can be found in the test evaluation sources listed above.

Test Development Procedures

The level of effort and technical expertise required to develop and maintain their own cognitive abilities test may be prohibitive for many organizations. Test development activities include development of test specifications and item pools, technical and sensitivity reviews, creation of test administration procedures/instructions, trying out items, analyzing item-level data, assembling the test, and preparing test documentation (Childs, Baughman, & Keil, 1997).

Develop Test Specifications

Test specifications are required to guide test development activities. See Russell and Peterson (2006) in this volume. Because reading of technical information (e.g., manuals) is required for electricians, a verbal comprehension test for applicants for the apprentice electronics job could be used. Test specifications include an operational definition of the construct to be measured,

content taxonomy, item reading level, item difficulty level, item format, item homogeneity, and number of items.

Construct definition. A clear operational definition of the construct to be measured must exist prior to beginning test content development. The construct definition should include a label, a brief definition, and information that distinguished the construct (e.g., verbal comprehension) from related constructs (e.g., verbal reasoning, word knowledge).

Content taxonomy. After the construct has been specified, the particular content used to measure the construct must be specified. The content for a verbal comprehension test can be described by grammar, word knowledge, making inferences, finding facts, seeing relationships, and identifying the main idea of the text.

Reading and difficulty level. The appropriate reading and item content difficulty level should be identified prior to test development during the job analysis. For cognitive ability tests being used in employment settings, appropriate reading and item difficulty level depend on the job requirements, the ability level of the intended applicant population, and the ability of the items to differentiate among applicants' ability level. For example, a 10^{th} grade difficulty level would be appropriate if the results of the job analysis support it and the target population was high school graduates with little or no college. A higher level would be

appropriate if job requirements were more demanding and the target population was college graduates. For tests not requiring verbal ability (e.g., numerical memory, spatial reasoning) and intended for the general population, a lower reading level might be appropriate. If too high a reading level is used, differences in performance on test content might be obscured by differences in reading skill.

Item format. Although multiple-choice formats are widely-used, cognitive ability test items may take other forms, including essay, true/false, and short answer. The Educational Testing Service, for example, recently added essay questions to the SAT because several colleges and universities wanted an indicator of applicants' ability to express themselves verbally. Regardless of which format is used, the item content should be representative of the cognitive processes the test is intended to measure (construct validity).

For multiple-choice tests, items are composed of a stem and response alternatives. The correct response is called the keyed response; all other response alternatives are distracters. Childs et al. (1997) noted several issues that should be considered when developing multiple-choice test items.

Items should be well-organized and clearly written, using familiar words and brief, direct statements. Item stems should be complete and provide enough information so the question is clearly stated. It should not be necessary for the examinee to have to read the response alternatives to understand the test question.

The response alternatives should be logically and grammatically consistent with both the item stem and the other response alternatives. Each item should have a single key. That is, there should be only one alternative that is clearly superior to the others. None of the distracters should be close enough in meaning to the keyed response that they could be defended justifiably as the correct response. Although the distracters should not be close enough in meaning to the item key that they can be justifiably defended as correct, they should be plausible enough to be effective. Well-written distracters should appear on the surface to be plausible in order to attract examinees that only superficially read the alternatives or do not know the content sufficiently. The response alternatives also should be written to be mutually exclusive. That is, no response alternative should logically contain another. Finally, all items should be independent from one another. Neither the item stem nor the response alternatives should include information that suggests the correct answer to another item. An example might be an item such as "Ohm's Law defines the relations between what variables?" followed by another question that includes information about the relations between power, voltage, current, and resistance.

Item homogeneity. Item homogeneity is inversely related to the breadth of the ability being measured. Tests designed to measure narrowly defined cognitive abilities such as verbal comprehension will be comprised of very similar items. In

contrast, a test of a more broadly defined concept such as verbal ability may contain items with varied content (verbal comprehension, verbal reasoning, verbal working memory, written expression). Regardless of the specificity or breadth of the ability being measured, the items 1) should be representative of the ability they are intended to measure and not measure other abilities as well and 2) should not contain content that may confound measurement of the targeted ability.

Number of items. The number of items for a test will be a function of the breadth of content, item format, and response format. For example, we may decide that in addition to questions focusing on comprehension, our test of verbal comprehension should also include items that measure grammar and word knowledge. It should be decided how many items of each type are desired prior to beginning to write test items. Detailed specification of the number and types of items for a test facilitates test construction and helps ensure comparability across forms, when multiple forms of a test are to be developed.

Figure 1 provides an example passage and questions that could be used to assess verbal comprehension for applicants to an entry-level electrician job training program. Note that he content of the passage, about Thomas Edison's many inventions, is of potential interest to applicants and is likely to increase applicant acceptance of the test (i.e., face validity). Verbal comprehension passages need not be so lengthy. Items from the ASVAB Paragraph

Comprehension subtest consist of short passages each followed by a single question. Figures 2 and 3 provide example test questions that could be used to assess math and spatial perception for applicants to the entry-level electrician job training program.

Detailed item writing guidelines are available elsewhere. See for example, Millman and Greene (1989) and Roid and Haladyna (1982).

Conducting Technical Reviews

Technical reviews are formal procedures in which subject matter experts (SMEs) and testing experts review test materials prior to field testing. The purpose of the SME review is to ensure the technical accuracy of the test items. For example, SMEs would determine that the item key is correct and that the item distracters are incorrect and are not ambiguous or misleading. The purpose of the test expert review is to make sure that the test items follow the item development guidelines described earlier and that the item content reflects an appropriate level of the ability being assessed as was determined in the job analysis.

Conducting Sensitivity Reviews

Sensitivity reviews are formal procedures in which representatives from various demographic groups review test materials to ensure they do not contain

content that may be viewed as *potentially offensive*. As a rule, sensitivity reviews focus on three issues. These are whether 1) test materials include assumptions, stereotypic descriptions, or objectionable or demeaning characterizations of subgroups, 2) test items might give one subgroup an advantage over others, and 3) items contain content about potentially sensitive topics. Some example of content that might give one group an unfair advantage over another include topics typically more familiar to one sex group (e.g., fashion, sports trivia), activities that are more accessible to members of higher socioeconomic status groups (e.g., equestrian, golf, possession of a private pilot's certificate), and local or regional knowledge (local/regional geography, history, or customs). Examples of potentially sensitive topics include religion, sex, and politics. These topics should be avoided. Further, items should not include offensive terms or language.

Creation of Test Administration Procedures and Instructions

Standardization of test administration procedures and instructions is necessary to ensure that the test-taking experience is as similar as possible for all examinees. Standardization provides administrators a process to control potentially extraneous factors (e.g., physical testing conditions) that may affect examinees' performance and ensure that test scores are comparable across test administrations.

Test administration procedures should include detailed instructions regarding the need for specialized training or certification for test administrators, setting up the test room, handling test materials, timing the test, and scoring completed tests. Test administration instructions should clearly describe standards for the physical testing environment (e.g., light, noise, temperature) and include information about how administrators should respond to examinees' questions. Test administration guidelines may go so far as to provide answers to clarifying questions that are frequently asked by examinees.

Trying Out Test Administration Procedures and Test Items

It is important to try out test administration procedures and test items prior to operational administration. This is essential for newly-developed tests where there is little or no prior data. Pilot testing provides an opportunity to refine administration procedures such as timing. For example, if a 45-minute time limit has been set for our Verbal Comprehension test, but most examinees complete it in less than 30 minutes; we might want to reduce the time limit. Another example is if we were developing a speeded test, where item content is very easy and test performance is a function mostly of response speed. If the time limit is set too long and most participants are finishing all of the items, the test may not be achieving its objective of measuring response speed.

Sometimes in an on-going testing program in a routine test administration, it is desirable to administer new unscored items along with those that are scored. The new items can either be embedded in the test with the scored items or presented in a separate section. These new items do not contribute to examinees' scores. They are used solely to determine the item characteristics (reliability, validity, difficulty, discrimination) and may become candidates for inclusion in a subsequent form of the test.

Analyzing Item-Level Data

Pilot studies provide a valuable source of data to conduct statistical analyses of item-level characteristics. Item-level statistical analyses focus on determining the difficulty level and discrimination index of the test items and help to guide test construction (e.g., number of items needed to reach a target reliability level).

Item difficulty. Item difficulty usually is measured by the proportion of examinees that correctly answer an item (p). Determining the appropriate range and mix of item difficulties is a crucial step in test construction. Consider two extreme examples. Suppose a test is constructed where all examinees answer some very easy items correctly ($p = 1.00$) and all examinees answer a different set of very difficult items incorrectly ($p = 0.00$). Clearly, neither the very easy nor the

very difficult test items are informative as they do not provide information that allows us to discriminate (i.e., makes distinctions) among the examinees on the targeted ability or to predict some external criterion (e.g., training or job performance). A test without variance cannot predict any criterion.

Items that provide the best discrimination among examinees, from a psychometric standpoint, are those that are answered correctly by about half of those taking the test. If the goal were to maximize the number of distinctions among the examinees with respect to the targeted ability, the best approach would be to develop a test where the average proportion of correct responses is about .50 across a range of item difficulty levels. This is frequently accomplished by using items with difficulty values ranging from .20 to .80. However, if the goal were to increase the probability of screening for a particular ability level, that might require the use of a minimum qualifying (or cutting) score, the best strategy would be to select items such that the average proportion of correct responses was equal to the selection ratio (i.e., number of openings/number of applicants). For example, if there were 300 applicants for the apprentice electrician job training program and 75 openings, the best test would be one that identified the top 75 applicants. For this example, the best test would be one where the average *p* value for the test items was .25. The difficulty level would be such that only about 25% of the examinees would obtain a qualifying score.

Item discrimination. Item discrimination typically is measured by the point-biserial correlation coefficient (Crocker & Algina, 1986). The point-biserial correlation indicates the degree to which performance on the test item is related to overall test performance. A test item discriminates between good and poor overall test performance to the extent that examinees who answer an item correctly also do well on the entire test. One common guideline is to retain items where the point-biserial correlation for the correct response (key) is .20 or greater. The point-biserial correlations for item distracters (incorrect options) should be negative. That is, those choosing the incorrect option should also obtain a lower score on the test as a whole. Some test constructors prefer to use the biserial correlation rather than the point-biserial correlation. The biserial correlation is not a Pearson correlation, but can be tested for significance and does not have the limitation of the point-biserial correlation with respect to extremely difficult or extremely easy items. A useful guideline is to retain items with a biserial correlation of .30 or greater.

Item bias. When examining item-level data, it is also informative to compute indices of item bias (Holland & Wainer, 1993). For example, indices of differential item functioning (DIF) can provide information about whether individuals from different subgroups (e.g., sex, ethnic, or socioeconomic groups), when matched on overall test performance, perform differently on individual test

items. Analyses based on item response theory (IRT) also can provide information about item difficulty, discrimination, and potential bias. Items that show signs of being problematic, statistically speaking, should be reviewed carefully to determine possible structural and content-related sources that may contribute to the problem (e.g., distracters that are too close to the item key, unintentional cues to the correct answer).

Item banking. Commercial item banking programs are available. These programs allow the test constructor to keep records about specific test items on a computer and manipulate test content. Also, many "spreadsheet" type programs can be used for item banking as they allow both numeric (item difficulty, item discrimination, sample size, etc.) and character (stems, keyed responses, distracters) data. Additionally, these programs allow sorting and filtering of data as well as statistical computations.

Assembling the Test

Several sources of information enter into the determination of the final set of items that will appear on a test. The results of the item-level analyses play a major role in the decision to include or exclude items. For example, item difficulty should be considered in order to achieve an appropriate range of item difficulty and overall test difficulty. In addition, point-biserial (or biserial)

correlations should be considered to attain internal consistency. Further, test specifications should be consulted that stipulate the number, type, and content of test items. Other sources of information that should be considered when assembling the tests include the similarity of item content to other items. Items should not be included that are redundant or provide information that may cue examinees regarding the correct answer to other items.

Although the results of the item analyses from the pilot test data provide much useful information when initially assembling the test, it is essential to conduct additional item-level analyses once the test has been operationally implemented. Test items may perform differently in an operational setting than they did during test development. There are several reasons why this may occur, including differences in administration procedures, examinee ability level and motivation, and position of the test items in the test. Therefore it is good practice to conduct additional item-level analyses once a test has been administered operationally.

Preparing Test Documentation

Test documentation materials should include a detailed summary of the pilot test results. For instance, pilot test results may include information about areas of needed improvement (e.g., test administration procedures) and the results

of such changes. This summary also should specify the number of test items that meet the predefined standards for inclusion in the operational test. Other useful information that should be provided includes distributions of surviving items sorted by difficulty level and point-biserial (or biserial) correlations.

SUMMARY

Individuals with high job performance are important to organizations. Campbell, Gasser, and Oswald (1996) examined the productivity of individuals with high and low job performance. Using a conservative method, they estimated that the top one percent of workers produces a return 3.29 times as great as the lowest one percent of workers. Further, depending on the variability of job performance, Campbell et al. estimated that the value may range from 3 to 10 times the return. Job performance makes a difference and the best predictor of job performance is cognitive ability (Schmidt & Hunter, 1998)!

Cognitive ability has a long history in occupational psychology. Measures of cognitive ability have been used to assess human capability and job qualification for literally centuries. The decision to select or develop a test begins with job analysis and the chapter by Anderson, Harvey, Baranowski, & Morath (2006) in this volume provides a handy reference.

Throughout this chapter, we have stressed the complexity of the decision process in purchasing or developing a cognitive ability test. Information about a test should include 1) test theoretical basis, normative sample, and test development procedures, 2) estimates of psychometric characteristics such as reliability, validity, and test bias, 3) administration procedures including materials, instructions, reasonable accommodation for applicants with disabilities, and special training required, 4) normative data, expectancy charts, or suggested cut scores, 5) scoring options and the qualifications for scoring, 6) existing or planned research/refinement of the test, 7) testing time requirements, 8) qualifications of the test developers, and 9) total costs including life-cycle costs.

A checklist is provided in Figure 4. Use the first line to enter the name of the test and write comments about the status of the "item." Multiple checklists can be used for comparisons.

The responsibility for defending the use of a cognitive ability test ultimately falls on the organization that uses it. Thus, it is crucial that the theoretical basis of the test be well understood, that it has acceptable psychometric properties, and that the test be administered, scored, and interpreted in an appropriate manner. Decisions made, even in part, on the basis of applicants' performance on cognitive tests have real-world consequences that affect individuals' lives (e.g., entrance into a training or educational program,

employment, promotion). Those who are screened out for entrance into training, hiring, or promotion based on their performance on a cognitive ability test may have little recourse. Further, they may not have the opportunity to demonstrate other competencies (e.g., job knowledge, motivation, skills) that may compensate for low cognitive test scores. Their only recourse may be in the courts.

REFERENCES

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (Joint Committee) (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anderson, L. A., Harvey, J., Baranowski, L., & Morath, R. (2006). Overview and description of inductive methods of job analysis. In D. L. Whetzel and G. R. Wheaton (Eds.) *Applied Measurement: Industrial Psychology in Human Resources Management.* Mahwah, NJ: Erlbaum.

Arvey, R. D. (1986). General ability in employment: A discussion. *Journal of Vocational Behavior, 29,* 415-420.

Campbell, J. P., Gasser, M. B., & Oswald, F. L. (1996). The substantive nature of job performance variability. In K. R. Murphy (Ed.). *Individual differences and behavior in organizations* (pp. 258-299). San Francisco: Jossey-Bass.

Carretta, T. R. (1997). Group differences on US Air Force pilot selection tests. *International Journal of Selection and Assessment, 5,* 115-127.

Carretta, T. R., & Ree, M. J. (1993). Basic Attributes Test (BAT): Psychometric equating of a computer-based test. *The International Journal of Aviation Psychology, 3,* 189-201.

Carretta, T. R., & Ree, M. J. (1995). Near identity of cognitive structure in sex

and ethnic groups. *Personality and Individual Differences, 19,* 149-155.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Childs, R. A., Baughman, W. A., & Keil, C. T., Jr. (1997) Tests of cognitive ability, in D. L. Whetzel and G. R. Wheaton, *Applied measurement methods in industrial psychology,* (pp 143-183). Palo Alto, CA: Davies-Black Publishing.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Ft. Worth, TX: Harcourt Brace Jovanovich.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16, 297-334.*

DeFries, J. C., Vandenberg, S. G., McClearn, G. E., Kuse, A. R., Wilson, J. R., Ashton, G. C., & Johnson, R. C. (1974). Near identity of cognitive structure in two ethnic groups. *Science, 183,* 338-339.

Equal Employment Opportunity Commission (1978). *Uniform guidelines on employee selection procedures.* Title 29 - Labor, Part 1607. *National Archives and Records Administration code of federal regulations.* Washington, DC: U.S. Government Printing Office.

Eyde, L. E., Robertson, G. J., Krug, S. E., Moreland, K. L., Robertson, A. G., Shawan, C. M., Harrison, P. L., Porch, B. E., Hammer, A. L., & Primoff,

E. S. (1993). *Responsible test use: Case studies for assessing human behavior.* Washington, DC: American Psychological Association.

Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance: The description of human tasks*. Orlando, FL: Academic Press.

Gael, S. (1988). *The job analysis handbook for business, industry, and government.* Vols. 1 and 2. New York: Wiley.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences.* London: MacMillan.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Goleman, D. (1995). *Emotional intelligence.* NY: Bantam Books.

Gottfredson, L. S. (1986). Societal consequences of the *g* factor in employment. *Journal of Vocational Behavior, 29,* 379-410.

Gottfredson, L. S. (1997). Why *g* matters: The complexity of everyday life. *Intelligence, 24,* 79-132.

Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin, 53,* 267-293.

Guilford, J. P. (1959). Three faces of intellect. *American Psychologist,* 14, 469-479.

Gustafsson, J. E. (1980, April). *Testing hierarchical models of ability organization through covariance models.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence,* **8**, 179-203.

Gustafsson, J. E. (1988). Hierarchical models of individual differences in cognitive abilities. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence,* (Vol. 4, pp. 35-71). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (Eds.) (2993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Holzinger, K. J., & Harmon H. H. (1938). Comparison of two factorial analyses. *Psychometrika, 3,* 45-60.

Humphreys, L. G., & Taber, T. (1973). Ability factors as a function of advantaged and disadvantaged groups. *Journal of Educational Measurement, 10,* 107-115.

Hunter, J. E. (1980). *Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB).* Washington, DC: U.S. Department of Labor, Employment Service.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72-98.

Hunter, J. E., & Schmidt, F. L. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86,* 721-735.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis (2nd ed.).* Thousand Oaks, CA: Sage.

Jensen, A. R. (1980). *Bias in mental testing.* New York: The Free Press.

Jensen, A. R. (1993). Spearman's *g*: Links between psychometrics and biology, *Annals of the New York Academy of Sciences, 702,* 103-129.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CN: Praeger.

Keyser, D. J., & Sweetland, R. C. (Eds.). (1997). *Test critiques* (Vol. 11). Austin, TX: Pro-ED.

Mayer, J. D., Salovey, P. & Caruso, D.R. (2002). *Mayer Salovey Caruso Emotional Intelligence Test.* North Tonawanda, NY: MHS.

McArdle, J. J. (1996). Current directions in structural factor analysis. *Current Directions in Psychological Science, 5,* 11-18.

McClelland, D. C. (1993). Intelligence is not the best predictor of job performance. *Current Directions in Psychological Science, 2,* 5-6.

McCormick, E. J. (1976). Job and task analysis. In M. D. Dunnett (Ed.), *Handbook of industrial and organizational psychology* (pp. 651-696). Chicago: Rand McNally.

McCormick, E. J. (1979). *Job analysis: Methods and applications.* New York: AMACOM.

Michael, W. B. (1949). Factor analyses of tests and criteria: A comparative study of two AAF pilot populations. *Psychological Monographs, 63*, 55-84.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 335-366). Washington, DC: American Council on Education.

Murphy, L. L., Plake, B. S., Impara, J. C., & Spies, R. A. (Eds.). (2002). *Tests in print IV: An index to tests, test reviews, and the literature on specific tests.* Lincoln, NE: University of Nebraska Press.

Olea, M, & Ree, M. J. (1994) Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology, 79,* 845-851.

Peiró, J. M., & Munduate, L. (1994). Work and organizational psychology in Spain. *Applied Psychology: An international review,* **43**, 231-274.

Plake, B. S., Impara, J. C., & Spies, R. A. (Eds.). (1993). *The fifteenth mental measurements yearbook.* Lincoln, NE: University of Nebraska Press.

Rabbitt, P., Banerji, N., and Szymanski, A. (1989). Space fortress as an IQ test? Predictions of learning and of practiced performance in a complex interactive video-game. *Acta Psychologica, 71,* 243-257.

Ree, M. J., & Carretta (1995). Group differences in aptitude factor structure on the ASVAB. *Educational and Psychological Measurement,* **55,** 268-277.

Ree, M. J., & Carretta, T. R. (1996). Central role of *g* in military pilot selection. *The International Journal of Aviation Psychology, 6,* 111-123.

Ree, M. J., & Carretta, T. R. (1998). General cognitive ability and occupational performance. In C. L. Cooper & I. T. Robertson (Eds.). *International review of industrial and organizational psychology, 13,* (159-184), Chichester, UK: John Wiley.

Ree, M. J., Carretta, T. R., & Doub, T. (1998/1999). A Test of Three Models of the Role of g and Prior Job Knowledge in the Acquisition of Subsequent Job Knowledge. *Training Research Journal, 4,* 1-16.

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulae. *Journal of Applied Psychology, 79,* 298-301.

Ree, M. J., Carretta, T. R., & Teachout, M. S. (1995) Role of ability, and prior knowledge in complex training performance. *Journal of Applied Psychology, 80,* 721-730.

Ree, M. J., & Earles, J. A. (1993). g is to psychology what carbon is to chemistry: A reply to Sternberg and Wagner, McClelland, and Calfee. *Current Directions in Psychological Science, 2,* 11-12.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance; Not much more than *g*. *Journal of Applied Psychology, 79,* 518-524

Ree, M. J., & Wegner, T. G. (1990). Correcting differences in answer sheets for the 1980 Armed Services Vocational Aptitude Battery reference population. *Military Psychology, 2,* 157-169.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing.* Orlando, FL: Academic Press.

Russell, T. L., & Peterson, N. G. (2006). Measurement plans and specifications. In D. L. Whetzel and G. R. Wheaton (Eds.) *Applied Measurement: Industrial Psychology in Human Resources Management.* Mahwah, NJ: Erbium.

Schmidt, F. L., & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general mental ability, and job knowledge. *Current Directions in Psychological Science, 2,* 8-9.

Schmidt, F. L., & Hunter, J. E. (1982). Two pitfalls in assessing fairness of selection tests using the regression model. *Personnel Psychology, 35,* 601-607.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Schmidt, F. L., & Hunter, J. E. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86,* 162-173.

Schulte, M. J., Ree, M. J., & Carretta, T. R. (2004). Emotional intelligence: Not much more than *g* and personality. *Personality and Individual Differences, 37,* 1059-1068.

Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures* (4th edition). Bowling Green, OH: Society for Industrial and Organizational Psychology.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: MacMillan.

Spearman, C. (1930). "G" and after- A school to end schools. In C. Murchison (Ed.), *Psychologies of 1930* (pp. 339-366), Worchester, MA: Clark University Press.

Spearman, C. (1937). *Psychology down the ages, volume II.* London: MacMillan.

Spearman, C. (1938). Thurstone's work reworked. *Journal of Educational Psychology, 39,* 1-16.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R. J., & Wagner, R. K. (1993). The *g*-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science, 2,* 1-5.

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score Reliability* (pp. 3-30). Thousand Oaks, CA: Sage.

Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior, 29,* 322-339.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs No. 1,* Chicago: University of Chicago Press.

Thurstone, L. L. & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs No. 2,* Chicago: University of Chicago Press.

Vernon, P. E. (1950). *The structure of human abilities*. New York: Wiley.

Vernon, P. E. (1969). *Intelligence and cultural environment*. London: Methuen.

Zhang, H. (1988). Psychological measurement in China. *International Journal of Psychology, 23,* 101-117.

Thomas Edison took out his first patent when he was 21 years old. It was for an electronic vote counter, which he intended for use in the United States House of Representatives. Although the machine worked perfectly, Congress would not buy it. The congressmen did not want the vote counting to be done too quickly. The roll call vote often was used to delay the voting process. Political groups relied on these delays to influence and change the opinions of their colleagues. Edison learned a valuable lesson from this experience; that is, "First be sure a thing is wanted or needed, then go ahead."

Born in 1847, Edison was the 7th and last child of Samuel Edison, Jr. and Nancy Elliot Edison. At an early age, he developed hearing problems that may have motivated him in the development of several of his inventions. To compensate for his deafness, Edison became an avid reader. Although Edison was inquisitive and imaginative, he had difficulty in school due to his hearing problems and only attended a total of 434 days over a five year period.

Edison created the first industrial laboratory in Menlo Park, NJ. At age 29, Edison began work on the carbon transmitter, which ultimately helped make Alexander Graham Bell's "articulating" telephone audible enough for practical use. In 1879, disappointed that Bell had beaten him in the race to patent the first authentic transmission of the human voice, Edison invented the first commercially

practical incandescent electric light bulb. Edison tested over 3,000 filaments before he came up with his particular version of a practical light bulb.

Many of Edison's inventions were in response to specific demands for new products or for improvements. However, he also had a gift for exploring unexpected direction when they were presented. Such was the case with the phonograph. The telephone was considered to be a variation of acoustic telegraphy. As with the telegraph, Edison was trying to develop a method to transcribe the signals as they were received. The recorded voice would then be retransmitted as a telegraph message. (The telephone was not yet conceived of as a general purpose method for person-to-person communication). In 1877, Edison used a stylus-tipped carbon transmitter to make impressions on a strip of paraffin-coated paper. To Edison's surprise, the barely visible indentations produced a vague reproduction of sound when the strip of paper was pulled back beneath the stylus. Edison subsequently replaced the paraffin-covered paper with a cylinder wrapped in tinfoil. The devise was universally acclaimed and Edison became known as the "Wizard of Menlo Park." It would be another decade however, till the phonograph moved from the laboratory to become a commercial product.

Perhaps Edison's greatest invention, however, was a practical and complete model for a standardized centralized electrical power system and its supplementary components. This revolutionary breakthrough influenced the

design, development, and success of all later power plants. Edison's design featured a unique transformer-controlled three-wire feeder grid. It was the first design to guarantee that electrical energy could economically power and light small, medium, and large communities worldwide. Despite its importance, the significance of this invention has largely been ignored and forgotten.

At the time of his death at age 84, Edison either singly or jointly had patented 1,093 inventions, including the incandescent light bulb, alkaline storage battery, phonograph, and motion picture projector. He also improved on the original design of other inventions such as the stock ticker, telegraph, and telephone. He believed in hard work, sometimes working 20 hours a day. This strong work ethic is reflected in a quote attributed to him that "Genius is one percent inspiration and 99 percent perspiration."

1. According to the passage, Congress decided not to purchase Edison's electronic vote counter because

A. it was too expensive to implement

B. of potential errors in vote counting

C. they preferred the roll call vote

D. electronic voting would lead to delays

2. According to the passage, Edison's work on the carbon transmitter contributed to the development of the:

A. stock ticker

B. telegraph key

C. feeder grid

D. articulating telephone

3. According to the passage, the phonograph was:

A. a response to demand from the entertainment industry

B. developed to record voice messages from telephones

C. preceded by the development of the telephone

D. an immediate commercial and financial success

4. According to the passage, Edison originally recorded sound on a:

A. strip of paraffin covered paper

B. paraffin covered paper cylinder

C. cylinder covered with tinfoil

D. solid wax covered cylinder

5. According to the passage, Edison:

A. was known as the "Wizard of Wall Street" due to his shrewd knack for investments

B. held patents in diverse areas, including the light bulb, phonograph, and automobile

C. was credited with creating the first industrial laboratory in Menlo Park, NJ

D. attended Princeton University where he earned a master's degree in science

6. According to the passage, Edison's greatest invention was the:

A. affordable incandescent light bulb

B. portable electric powered phonograph

C. first practical articulating telephone

D. model for a centralized power system

1. An electrician doing the wiring for a building estimates that 1,600 feet of electrical cable will be needed. Four spools contain 1,000 feet of cable. How many spools should the electrician purchase?

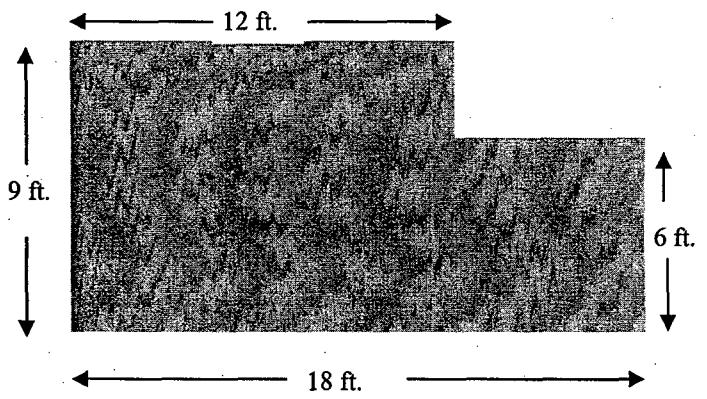
A. 2

B. 4

C. 6

D. 7

2. A contractor wants to purchase electrical insulating material for the area shown in the figure below. How many square feet of insulating material are needed to cover the entire area?



A. 108

B. 144

C. 162

D. 216

3. What is the volume of a cylinder designed to hold electrical equipment that is 8 inches tall and has a 1 inch radius? (Use $\pi = 3.14$)

A. 24.00 cubic inches

B. 25.12 cubic inches

C. 50.24 cubic inches

D. 78.88 cubic inches

4. The reciprocal of 10 is:

A. 0.05

B. 0.01

C. 0.10

D. 1.00

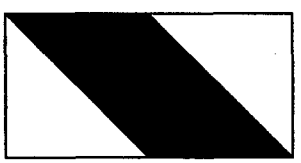
The following test items are designed to measure your ability to solve spatial problems. For each problem, you will be given three rules that will determine how four cubes are to be combined to form a larger figure. The cubes are grouped into two sets as follows:

Set 1 Set 2

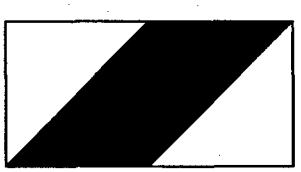
Figure A Figure B Figure C Figure D

Consider the following example:

Rule #1: "Figure B precedes Figure A" yields



Rule #2: "Figure D does not follow Figure C" yields

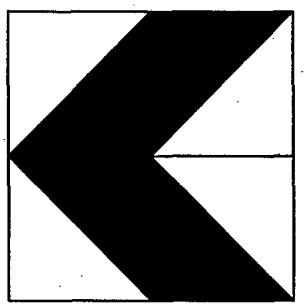


Rule #3: "Set 1 is below Set 2" yields

Set 2

Set 1

Combining the results of all three rules to create a larger figure yields:



Use the following figures to solve all of the remaining problems:

Set 1 Set 2

Figure A Figure B Figure C Figure D

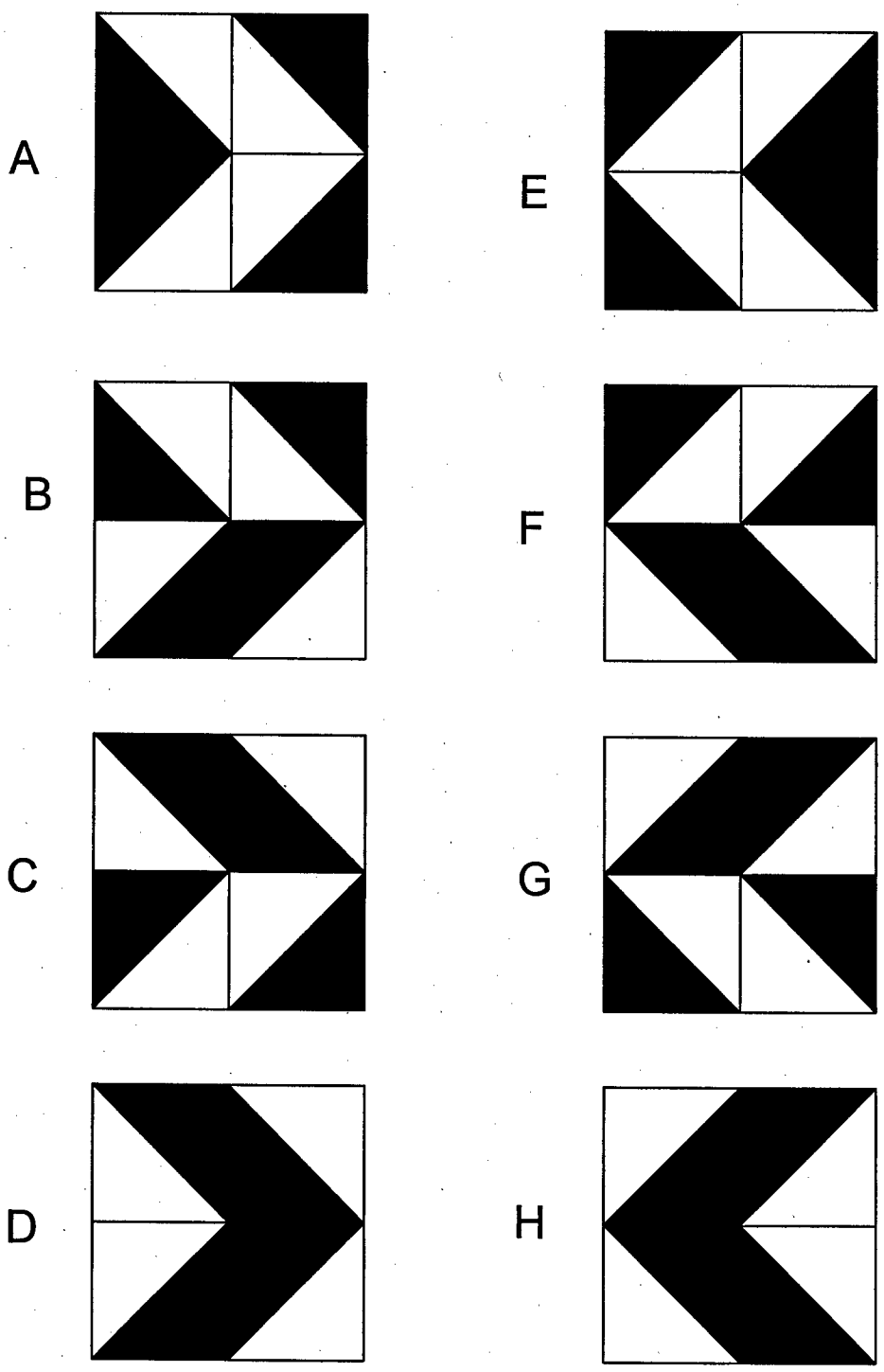
1. Use the following three rules to create a larger figure:

Rule #1: Set 2 is not below Set 1

Rule #2: Figure B does not follow Figure A

Rule #3: Figure C precedes Figure D

Choose the correct solution from the following alternatives.


A
B
C
D
E
F
G
H

CHECKLIST FOR _____

Item	Comment
1. Test theoretical basis	_____
2. Normative sample	_____
3. Test development procedures	_____
4. Reliability	_____
5. Validity	_____
6. Test bias	_____
7. Administration procedures	_____
Materials,	_____
Instructions	_____
Reasonable accommodation	_____
Special training required	_____
8. Normative data	_____
Expectancy charts	_____
Suggested cut scores	_____
9. Scoring options and the qualifications for scoring	_____
10. Existing or planned research/refinement of the test	_____
11. Testing time requirements	_____
12. Qualifications of the test developers	_____
13. Total costs including life-cycle costs	_____

[1] The views expressed are those of the authors and not necessarily those of the United States Government, the Department of Defense, or the United States Air Force.