Verify - Numerical Ability

Technical Manual

Contents

List of Tables

Chapter 1: Introduction

Specifying the Measurement Taxonomy for the Verify Tests

Historical Uses and Purposes of Cognitive Ability Tests

When compared with personality inventories, biodata, situational judgment, and structured interviews, cognitive ability is consistently the best predictor of performance across job levels and job types (Hunter, 1983; Murphy & Shiarella, 1997; Schmidt & Hunter, 1998; Wagner, 1997). Applications of ability testing can be divided historically into two general approaches: the use of general intelligence tests and the use of specialized ability tests. The widespread use of general cognitive ability tests for personnel selection and placement in industry can be traced in large part to the use of ability tests developed for selection and classification of U.S. military personnel during World Wars I and II (e.g. Salas, DeRouin, & Gade, 2007). More recently, the predictive validity of measures of general mental ability in occupational settings has been investigated using meta-analytic methods (Hunter & Hunter, 1984). These meta-analytic studies of ability-performance relations provided empirical evidence indicating the predictive validity of general cognitive ability measures for job performance across a wide range of jobs. Results of this line of research have led to general agreement that a substantial relationship exists between general cognitive ability and job performance (Hunter & Hunter, 1984; Kanfer, Ackerman, Murtha, & Goff, 1995; Ree & Earles, 1992).

However, based on the historical success of tailored ability batteries for predicting success in specific military roles (e.g., pilot, navigator) and continued success of using this approach in organizations, differentiated abilities tests (e.g., tests of verbal abilities, numerical reasoning, spatial ability) have also been adopted as a valid and efficient method of assessing candidates for job roles.

Structure and Organization of Intelligence

The history of human intelligence is diverse, from early theories that measured individual differences in cognitive abilities by psychophysical assessment (e.g., Galton, 1928) to contemporary theories that posited that intelligence is conceptualized as process, personality, interests, and knowledge (e.g., Ackerman, 1996). The theoretical evolution of human intelligence has guided research and practice over the past century in conceptualizing the importance of cognitive ability in basic and applied psychological issues.

The dimensionality of cognitive ability has been debated by psychologists and other scientists for the better part of a century (Schmitt & Chan, 1998). From Spearman's (1927) unidimensional theory of intellect (g) to Guilford's (1967) structure of intellect theory with 150 unique abilities, the granularity of how cognitive ability is defined varies greatly. Human abilities theorists have conceptualized the nature of cognitive abilities in a variety of ways (e.g., Cattell, 1963; Spearman, 1904). Spearman (1904) supported a unitary construct of intelligence and specified the two-factor model of intelligence, consisting of g, or general intelligence and s, or specific variance associated with a particular test. He concluded that ability-performance correlations showed hierarchical order and that sitting atop this hierarchy was a general intelligence factor. Spearman's two-factor theory of intelligence laid the groundwork for future measures of intelligence designed to elicit g. Such measures include Raven's Progressive Matrices (Raven, Court, & Raven, 1977, a nonverbal reasoning test) and Cattell and Cattell's (1960) attempt to create a 'culture-free' test of intelligence known as the Culture Fair Intelligence Test.

Thurstone (1938) was among the first to propose the notion of primary mental abilities. His group factors approach stood in contrast to Spearman's conceptualization of a unitary construct of intelligence. The seven factors included in Thurstone's conceptualization of abilities included: Verbal, Reasoning, Number, Spatial, Perceptual Speed, Memory, and Word Fluency. In combination, primary abilities reproduce a variety of intellectual functioning. His theorizing led to the development of the Primary Mental Abilities tests.

An alternative conceptualization of intelligence was proposed by Cattell (1963), who proposed an incomplete hierarchy (no '*g*') and argues that intellectual processes are organized into broad second order factors. Two of these factors, fluid intelligence (*Gf*) and crystallized intelligence (*Gc*), are most frequently associated with general intellectual functioning. *Gf* is postulated to relate to intelligence derived from neural-physiological factors of intellect while *Gc* is acquired through experiential and educational means. Horn and Cattell's (1966) theoretical basis for fluid and crystallized intelligence is that development influences the distinction between *Gf* and *Gc*. That is, *Gf* tracks physiological development and sets limits on what an individual can achieve in terms of *Gc*.

The fluid-crystallized intelligence spectrum encompasses a variety of primary abilities. At one end of the spectrum (that relating most closely with crystallized intelligence) are abilities that require information and practiced skill (e.g., knowledge tests, aspects of verbal ability). At the other end of the spectrum (that relating most closely with fluid intelligence) are abilities related to comprehending abstract or unfamiliar information and manipulating it to satisfy some requirement (e.g., inductive reasoning).

From a measurement perspective, tests designed to measure fluid and crystallized intelligence provide a broader test of intellectual functioning than test of general intelligence. Measures designed to elicit *Gf*, such as tests of verbal (e.g., verbal analogies), numerical-mathematical (e.g., problem solving), and spatial abilities (e.g., spatial orientation), are often administered along with measures designed to elicit *Gc* (e.g., vocabulary tests). Tests of fluid and crystallized abilities provide the opportunity for broader assessment of abilities for selecting high ability and high knowledge job candidates.

In 1993, in what is considered the largest factor analysis of cognitive ability testing data, Carroll derived the three-stratum theory of cognitive ability. Carroll's factor structure begins with a set of specific abilities (first stratum) that fall under eight broad factors (second stratum). All narrow and broad ability factors fall under a single general factor similar to Spearman's g. Carroll's theory is important due to the size of the factor analysis and how well Carroll's second stratum corresponds to the Cattell-Horn model. Both models have a crystallized and fluid intelligence factor, but Carroll's has six additional factors that include general memory and learning, broad visual perception, broad auditory perception, broad retrieval ability, broad cognitive speediness, and decision/reaction time/speed. Because of the strong empirical support for Carroll's model, the correspondence between both models, and the detailed theoretical grounding for both models, McGrew (1997) reconciled the two models to develop what is known as the Cattell-Horn-Carroll (CHC) theory. This theory is the most widely used theory in cognitive ability test development (Alfonso, Flanagan, & Radwan, 2003). The broad factors in the Cattell-Horn, Carroll, and CHC models are outlined in Table 1 [Carroll, 1993].

Table 1. Cattell-Horn, Carroll, and CHC Models of Cognitive Ability

Cattell-Horn	Carroll	CHC
• Fluid Intelligence • Crystallized Intelligence	• Fluid Intelligence • Crystallized Intelligence • General Memory and Learning • Broad Visual Perception • Broad Auditory Perception • Broad Retrieval Ability • Broad Cognitive Speediness • Decision/Reaction Time/Speed	• Fluid Intelligence/Reasoning • Crystallized Intelligence/Knowledge • General Knowledge • Visual-Spatial Abilities • Auditory Processing • Short-Term Memory • Long-Term Storage and Retrieval • Cognitive Processing Speed • Decision/ Reaction Time • Psychomotor Speed • Quantitative Knowledge • Reading/Writing • Psychomotor Abilities • Olfactory, Tactile, & Kinesthetic Abilities

The CHC model provided a strong theoretical backing to the Verify cognitive ability testing program of research and development. The model was used in conjunction with the O*NET Content Model (United States Department of Labor, n.d.) model to define the tests included in the range of Verify tests we offer. The O*NET model of cognitive ability is based upon the Fleishman taxonomy (Fleishman, Quaintance, & Broedling, 1984). The O*NET model was used because every job listed in O*NET is rated on the importance of each cognitive facet for successful performance of that job. This helps identify the most appropriate cognitive competencies to assess.

Numerical Ability and the Cognitive Ability Taxonomy

The Numerical Ability test is designed to measure the ability to solve problems involving numerical data by using the proper mathematical methods and the ability to interpret data presented in charts, graphs, and tables. Additionally, this test has been developed to assess the ability to understand percentages, fractions, decimals, proportions, basic geometry, basic probability, and the ability to compute solutions accurately using addition, subtraction, multiplication, and division. This test does not, however, assess numerical sequences, advanced geometry, statistics, measurement conversions (e.g., grams to pounds), or any concept that requires outside knowledge to solve (all required information is provided in question stems).

From a theoretical standpoint, numerical ability is an important component of cognitive ability (McGrew & Evans, 2004). It has been empirically demonstrated in numerous factor analyses of global cognitive ability measures that numerical ability is a distinct second-tier factor in hierarchical models with a first-tier global factor (e.g. Carretta & Ree, 1996). Among the 965 job titles listed on O*NET, 191 list Mathematical Reasoning as having an importance rating of at least 50%, and 163 list Numerical Facility as having an importance rating of at least 50%. Clearly, from both a theoretical and practical standpoint, numerical ability is an important facet of cognitive ability and plays an important role in successful performance for many jobs at a variety of levels. Numerical Ability falls under the Fluid Intelligence factor in the CHC model of intelligence, which provided a strong theoretical backing to our program of research and development.

Chapter 2: Test Materials and Use

Description of Test

The Numerical Ability Test measures the ability to recognize and attend to the relevant data in tables and charts, quickly and accurately use basic mathematical concepts to analyze data, to draw the appropriate conclusions based on mathematical analyses. It provides an indication of how a person will perform when they are asked to examine facts and figures and make sound decisions from data. This ability is commonly required to support work and decision-making in many different types of jobs at many levels.

Candidates are presented with 16 questions. They have 20 minutes to answer all of the questions. They are informed in the instructions that they will receive a score reduction if not all questions are answered.

The Verify Numerical Ability test was developed in U.S. English, but it is available in multiple languages. Please contact an account manager for details on language availability.

Intended Uses of the Numerical Ability Test

The Numerical Ability test can be administered individually, in combination with other ability tests, or in combination with other job-specific predictors of job performance. Normative data are available for making appropriate comparisons. Though tests are designed to be appropriate for a variety of job levels, a thorough job analysis is recommended to determine which tests are most appropriate for a given job.

The Numerical Ability test is appropriate for either proctored or unproctored administration. Due to the adaptive nature of the test administration, virtually every candidate will see a different set of questions, which alleviates the typical security concern with the use of cognitive ability tests in an unsupervised setting. Additional test security features, including candidate score verification, are fully outlined in the Verify User Guide.

Chapter 3: Foundations and Development

Test Development

The development of our Numerical Ability test began in 2013 and was led by a team of Industrial/Organizational Psychologists with extensive experience in selection testing, item[1] response theory (IRT), computer adaptive testing, and cognitive ability testing.

Test Properties

The following specifications were outlined:

- The test should contain at least 500 questions with stable IRT parameters in its pool. The questions should cover a range of difficulty with the greatest concentration at the median ability level.
- Questions should not cover any topic that might produce an advantage or disadvantage for a specific group.
  - Questions that demonstrate bias against members of a specific race, ethnicity, gender, or age group based on statistical analyses will be removed from the question pool.
  - No questions will be included that require previous factual knowledge (e.g. formulas, measurement conversions, historical facts).
- Questions will be presented one at a time. Questions will be in a multiple choice format with one and only one correct response option. There will be five response options for each question.
- The test will utilize our computer adaptive technology to administer questions adaptively.
- Candidates will be presented 16 questions and have 20 minutes to complete them.
  - Candidates that are unable to complete all 16 questions in the time permitted will receive a score reduction proportional to the number of questions left unanswered.[2]

Question Development

Because Numerical Ability is an adaptive test, a large bank of questions covering a broad range of difficulty was required. The development of the question bank was a multisource and multistage process that is outlined below.

The first source for questions was the Global Cognitive Index (GCI) Quantitative Ability test, and the second source was the Verify Numerical Reasoning test. The original content development for both tests involved the expert review of content gathered from several personnel selection testing programs, as well as question writing by Industrial/Organizational Psychologists within the organization. An extensive review of the legacy content was completed, involving two master's level and four doctoral level Industrial/Organizational Psychologists who rated each question on the following aspects:

- How job relevant is the question content?
- Is the reading level of the question appropriate for the difficulty of the question?
- Does the question have one and only one correct response option listed?
- Does the question have five response options?
- Does the question have charts, tables, graphs, or other graphics that will need to be generated, updated, or changed?

A combined question bank was used as a basis for our current Numerical Ability test. Questions for our Numerical Ability test were reviewed to decrease time required to complete and also underwent a decentering process to streamline localization. Questions that were deemed too U.S.-centric or were too long were revised or dropped. The bulk of the original questions, albeit revised and improved versions, were retained for use in our Numerical Ability test. This resulted in a total of 651 questions, utilizing 186 from the Verify Numerical Reasoning test and 465 from the GCI Quantitative Ability test.

Item Parameterization

As opposed to classical test theory scoring which is typically a tally of the total number of questions correct on a test, item response theory calculates scores based on unique information regarding an individual's performance on each question. Each question has parameters, such as difficulty, that contribute information regarding a candidate's ability level. Although the questions from the original Numerical Ability test that were kept in the item bank already had established parameters, question revisions and new test

[1] Note: "Item" and "question" are used interchangeably throughout this manual.
[2] The score reduction was not added until March 2014. See page 9 for additional details. Please contact SHL for information on tests included in jobs and solutions built prior to March 2014.

specifications (time and number of questions) warranted new parameter estimation. Therefore, all questions were re-trialed to collect data for item parameterization. The questions were administered using the same time and length specifications of the final product. Parameters for the questions were then derived from that data.

Test Construction

Our tests are constructed to balance fairness, efficiency, and precision. Candidates taking the Verify - Numerical Ability test have 20 minutes to complete 16 questions. These time and question quantity settings were established via trialing of multiple test-level timers. Based on the large portion (over 82%) of candidates completing 16 questions in 20 minutes, and the strong correlation between question difficulties from the untimed and timed test trials, the final configuration was determined.

Because the test is administered adaptively, an individual's theta (ability level) is estimated following each question he/she responds to. The standard error (SE) around that estimate is also calculated as a function of the item parameters for that question. As the candidate responds to more questions, the estimate of theta will become more accurate and the standard error will decrease. An SE of 0.45 was the precision goal for the legacy, variable-length, cognitive tests. Based on previous data, the 16-question length of the Numerical Ability test will yield precise ability estimates.

Score Reduction

The measurement precision associated with CAT scores has been widely cited as a key benefit of this approach to testing (Johnson & Weiss, 1980; Moreno & Segall, 1997; Embretson & Reise, 2000). Highly accurate scores can be achieved in less time and with fewer questions as compared to traditional, non-adaptive testing. This is especially critical when making decisions about candidates' qualifications and competencies during a personnel selection process.

Though CATs are capable of generating valid and reliable scores with fewer questions, candidates must complete enough questions to produce an accurate and reliable score. There are cases in which a less-than-optimal amount of information is available to produce reliable test scores. This is especially true when CATs have test-level timers. During the test development process, timers are set to ensure that most candidates can complete all or most of the questions on a test. However, some candidates may exit a CAT prematurely, and/or they may not attempt as many test questions as needed to produce a highly reliable score. These instances lead to test sessions with a less-than optimal degree of information about these candidates.

Prior to the introduction of the score reduction method, to ensure that reliable scores are produced on our Verify CATs, our strategy was to establish a minimum estimated reliability (maximum standard error) threshold, under which a candidate would receive an "invalid" test result if a minimum number of test questions was not completed. Because of the burden this method placed on test administrators in client organizations, we initiated a program of research to identify an alternative approach that yielded a valid score for all candidates regardless of how many questions they completed.

Many of the largest testing programs in the world are now administered as CAT. We investigated several of these programs to determine how those with test level timers dealt with examinees that do not complete all the questions in a test. The GMAT (Graduate Management Admissions Council [GMAC], 2012), ASVAB (Personnel Testing Division Defense Manpower Data Center, 2006), and Selection Instrument for Flight Training (SIFT; United States Army Recruiting Command, 2013) use CAT with a penalty for incomplete test sessions. [Other CAT programs investigated like the Graduate Record Exam, which no longer adapts at the question level, and the National Council Licensure Examinations (Registered nurse certification), which has a variable test length, differ too much from our Verify CATs to provide useful comparisons.]

Alternative Approaches for Scoring Incomplete Test Sessions

We investigated multiple alternative approaches to overcoming the challenge of candidates not completing a sufficient number of questions to produce a reliable score. We focused our investigation on a penalty-based approach. Because adaptive testing is not scored in a simple right/wrong format like linear or static testing, applying penalties is relatively complex (Segall, 1988). We highlight some key considerations below:

- In CAT, the fewer questions answered, the more biased the scores are toward the mean. We use an ability estimation algorithm that uses the population ability distribution as the basis for estimating ability. This means that before a candidate answers a single question, we assume that his or her ability level is at the mean because this is the densest part of the distribution. As a candidate answers more questions, the algorithm relies less and less on the population distribution and converges on a more reliable and precise estimate of the candidate's true ability level. Therefore, a score penalty should account for the number of questions completed and the corresponding information available about the candidate's estimated ability level.

- In addition to the bias toward the mean, fairness and construct measurement issues are appropriate to consider when comparing scores between people that completed differing numbers of test questions (e.g., all questions vs. a small number of questions). In a timed CAT, power (accuracy) and speed are both important measurement factors, so the highest possible scores can and should be achieved by individuals that can answer all of the questions on a test correctly. On the other hand, if no time limit were enforced, a skewed distribution of disproportionally high scores would result. As such, candidates who spend proportionally too much time completing a smaller number of questions in an effort to correctly respond to those questions should not receive a score equivalent to a candidate who correctly responds to a larger number of questions. This is consistent with the test design, which considers that cognitive processing speed as well as accuracy are both important determinants of the targeted cognitive ability constructs such as Deductive Reasoning, and thus are both likely to be important in prediction of job performance. Candidates who can accurately complete an entire timed test are more effectively demonstrating this construct. Therefore, enforcing on a timed cognitive CAT a score penalty which increases as fewer questions are completed appears to be in conceptual alignment with the measurement of these ability constructs.

The central criteria for the development and selection of a score penalty method included maintenance of criterion-related validity when applied to existing test data, a theoretical basis to ensure that scores accurately reflect the construct of cognitive ability, and broad applicability to all of our Verify CATs. We also wanted to ensure that no protected groups were adversely affected by the score penalty process. Determining the algorithm to apply to incomplete test sessions that met all of our theoretical, practical, logistical, and fairness requirements involved review of the literature, review of methods used in other large-scale CAT programs, and consultation with a measurement expert who oversees research and development for a large-scale educational testing program. We identified the proportional adjustment approach through our research and consultation. In this method, an algorithm is applied whereby candidates who do not complete a test are penalized proportional to the number of questions that they did not attempt. The resulting scores very closely approximate the score a candidate would receive if he or she had randomly guessed on all of the remaining questions in the test.

Upon identifying the proportional adjustment method as our recommended approach, we validated it with existing Verify test data. We applied the method to client test data from multiple Verify tests we had available and determined that it met all logistical and psychometric requirements. When recalculating validity coefficients using the proportional adjustment method, validities matched or exceeded those calculated using the previous method. These analyses are available upon request. The proportional adjustment method maintained criterion-related validity, is theoretically grounded, has an even score distribution, and does not penalize candidates based on question characteristics. As such, it was identified as the most effective score penalty approach based on this investigation.

Job Analysis

Overview of Consortium Research Settings, Samples, and Goals

Multiple organizations supported the criterion-related validation of our Numerical Ability test. These partner organizations also conducted job analyses to determine the importance of numerical ability to the positions in question. All of the job analyses included the following steps:

1. Completion of a Job Analysis Questionnaire (JAQ) by Subject Matter Experts (SMEs),
2. Data analysis and synthesis of findings, and
3. Recommendation of a set of tests based on the results of the job analyses for inclusion in a solution (a group of tests) for a criterion-related validity study.

Summary of Job Analysis

The process for determining the relevance of numerical ability for various positions occurred through a series of job analysis activities conducted with a variety of organizations in different industries. The goal of the project was to gather information about the present and future job requirements of these positions in order to identify the appropriateness of the Numerical Ability test. The breadth of the JAQs differed based on the organization's need, but for the sake of parsimony, only the findings as they relate to the development of this test are summarized.

Three samples provided JAQ ratings for the professional / individual contributor level and one sample provided ratings for the frontline manager role. Our research scientists distributed JAQs to SMEs the partner organization identified. Responses were not included in analyses if the SME lacked tenure, had too much missing data, had erratic response patterns (e.g. responding with all 5s), or did not consider themselves knowledgeable enough about the job.

The Verify Numerical Ability test was rationally linked to the competencies and work behavior dimensions in the JAQ. Table 2 contains the mean importance rating for the linked work behavior dimensions. The weighted average rating across the three samples is displayed. Ratings were made on a five-point scale of importance (1 being unimportant and 5 being critically important). As the table indicates, the relevant work behavior dimensions were rated as important for successful performance by respondents.

Table 2. Work Behavior Dimensions for Professional / Individual Contributor Roles

Work Behavior Dimension	Mean Rating for Professional / Individual Contributor Role	Mean Rating for Front Line Manager Role
Choosing and Applying Formulas and Complex Operations: involves applying mathematical formulas accurately, computing descriptive statistics, and processing data properly.	2.39	3.08
Math Problem Solving: involves identifying the correct arithmetical techniques to arrive at a solution and accurately calculating the numbers.	2.29	2.94
Math Basics: involves simple arithmetical calculation without the use of a calculator and the conversion among decimals, percentages, and fractions.	1.87	2.57

Chapter 4: Standardization, Scaling and Normative Reference Groups

Description of Scale and Question Types

All of the Numerical Ability questions are multiple choice. In these questions, candidates are presented with a question stimulus. Candidates are then asked to choose the correct answer. Every question has five response options. The Verify Numerical Ability test measures this competency using several types of questions:

Word Problems: In these questions, candidates are presented with a few sentences that provide a scenario that requires a numerical solution. Candidates must use the information provided to determine which numerical calculations are required and execute those calculations accurately.

SAMPLE:

Janis has an MBA and has been in school for 1/2 of her life. Carol is 30 years old and started school later, having only been in school for five years. If Janis is four years younger than Carol, how long has she been in school?

A. 13 years
B. 15 years
C. 18 years
D. 20 years
E. 26 years

The correct answer is **A**. Since we know Carol is 30 years old and that Janis is four years younger than Carol, Janis must be 26 years old. And, since we know that Janis has been in school for half her life, then she has been in school for 13 years.

Data Interpretation: These questions require candidates to correctly interpret data presented in tables, charts, or other figures. Some questions only require simple interpolation, whereas others also require some additional calculation.

SAMPLE:

Company Figures at the end of the Current Financial Year					
Figure	Company A	Company B	Company C	Company D	Company E
Turnover (£m*)	11.4	160.7	2.3	7.3	60.0
Profit (£m*)	9.2	-33.9	-1.9	3.0	9.3
Assets (£m*)	42.3	640.1	25.2	27.1	130.1
Debt (£m*)	9.7	341.0	1.7	8.2	30.8
Share Price (pence**)	108.0	172.1	6.0	11.3	48.7
No. of Shares (m*)	0.5	2.7	0.9	3.2	7.2
*m = millions **1 pence = 1/100 of a Pound £					

If someone were to sell 3,500 shares in Company A, how many shares of Company E could they buy with the proceeds from that sale?

A. 2,217
B. 5,260
C. 7,761
D. 15,782
E. 35,000

The correct answer is **C**. The table indicates that the Share Price for a share in Company A is 108.0 pence. This means that 3,500 shares costs 3,500 times that amount = 378,000 pence. A share in Company E costs 48.7 pence. The amount of pence on hand, 378,000 divided by the cost of one share, 48.7 pence, will equal the number of shares that can be purchased = 7,761.8. Because only whole shares can be purchased, the number is rounded to 7,761.

Normative Data and Decision Making

Test scores may be used in a variety of ways to make decisions about individuals. It is important for companies to determine how to make decisions based on a candidate's scores on any given test. Normative data help companies understand a candidate's relative standing by comparing his/her test scores to the scores of other candidates in the normative database. Companies may use this information to make decisions regarding the number of candidates who move on to the next phase of the hiring process. This is referred to as a normative approach because the normative data are used to estimate a passing score that will create efficiencies by minimizing the number of candidates proceeding to the next step, while providing enough candidates from which to fill job vacancies. As part of the test development process, all our tests are normed using a representative sample of test takers.

Normative information is embedded in score reports generated by completed tests. Normative data were computed on the basis of a database of candidate data from past assessment usage. The scores reported to hiring managers are in percentile, which compare how a specific candidate performed to the performance of the average candidate. In addition to comparing candidate scores to what we would expect from the average candidate, a percentile score considers the amount of variation around that average we would expect. This is the standard deviation. When a candidate is a standard deviation unit or more below the average, his or her percentile score begins to get very low. For example, if the average score on an assessment was 20 points and most candidates fell between 18 and 22 points, the standard deviation would be 2. If a candidate scored 16 points, he or she would be 2 standard deviation units below the average score and would receive a very low percentile score. If the average score was still 20 but most candidates fell between 16 and 24 the standard deviation would be 4. Then a candidate scoring 16 points would only be one standard deviation unit below the average and would have a higher percentile score even though his or her score and the average score are the same as in the first example. Any time we see a candidate with a score one or more standard deviation units below the average, he or she will have a very low percentile score. We use an approach whereby candidates who receive scores in the "Not Recommended" zone score below the 30th percentile.

Normative Reference Groups

The dataset used to calculate the normative information for the Numerical Ability test were candidate scores collected from individuals across a number of countries and job levels. The total number of participants in the sample was *N* = 6,927. The mean, standard deviation, skewness and kurtosis for our Numerical Ability test are presented in Table 3 below. The demographic breakdown of this sample is displayed in Table 4.

Table 3. Norms Statistics

N	*M*	*SD*	Skewness	Kurtosis
6, 927	-0.16	0.91	-0.40	-0.09

Table 4. Demographic Breakdown of Sample Used to Generate Norms

<table>
<tr><th>Demographic</th><th>%</th><th>Prefer Not to Answer</th></tr>
<tr><td colspan="3">Gender</td></tr>
<tr><td>Male</td><td>50%</td><td rowspan="2">15%</td></tr>
<tr><td>Female</td><td>35%</td></tr>
<tr><td colspan="3">Age</td></tr>
<tr><td><40 years</td><td>71%</td><td rowspan="2">18%</td></tr>
<tr><td>≥40 years</td><td>11%</td></tr>
<tr><td colspan="3">Ethnicity</td></tr>
<tr><td>Asian</td><td>24%</td><td rowspan="5">22%</td></tr>
<tr><td>Black or African-American</td><td>8%</td></tr>
<tr><td>Hispanic or Latino</td><td>2%</td></tr>
<tr><td>Two or More Races</td><td>1%</td></tr>
<tr><td>White</td><td>43%</td></tr>
<tr><td colspan="3">Country of Origin</td></tr>
<tr><td>United Kingdom</td><td>42%</td><td>0%</td></tr>
</table>

Demographic	%	Prefer Not to Answer
United States	9%	
Australia	9%	
South Africa	7%	
Other	33%	

Chapter 5: Reliability

Overview

The reliability of a test refers to the extent to which it is free from measurement error. Reliability allows one to interpret differences in test scores among individuals as true differences in the skill or trait being measured rather than something else. Methods for establishing reliability for "static" (non-adaptive) tests include internal consistency (e.g., coefficient α, KR-20), alternate forms, and test-retest reliability. Each of these can be used to provide estimates of reliability which, together with sample variance information can be used to compute the standard error of measurement (SEM). The first method (internal consistency) is not possible with Computer-Adaptive Tests (CATs) due to their adaptive nature (in other words, it is a requirement of internal consistency methods that the test content be exactly the same for each individual taking the test). Alternate forms reliability is not exactly appropriate for CAT-based tests either, since alternate forms assumes a finite number of different forms that are being compared, rather than the multitude of possible forms created through adaptive technology methods. Test-retest reliability is appropriate for CAT-based tests, but the study design required for yielding this estimate of reliability was out of scope for the current research project.

In classical test theory, the SEM is typically denoted as a 'fixed' value for any given test. That is, it implies that the error associated with any score on that test is the same for all scores. However, static tests tend to be more reliable for assessing candidates who are of average ability, and less reliable for those candidates who are of high or low ability. For CAT-based tests, measurement error is defined by the Standard Error of theta (SE). This error varies across the range of theta values depending on the shape of the test information function. Typically, though, CAT tests are designed to give reasonably low errors of estimate within a reasonably broad range of theta values. The *SE* can be used as a stopping rule in which the test administration engine can be programmed to end the test once a desired level of score precision has been reached.

Our computer adaptive tests are fixed in the number of questions administered. This is done to create a more consistent testing experience across candidates. Because the Numerical Ability test does not use standard error as a stopping rule, the reliability of the test is calculated in a different way. Using the test configuration of the actual test, including the question parameters of the entire question bank, test performance was simulated. Because IRT scoring is probability based, it is possible to simulate how a candidate of a given level of ability will perform on a CAT. The benefit of using simulated candidate scores is that the true score is a known value. For real candidates, the true score can only be estimated. A sample of 50,000 simulated candidates with normally distributed ability was generated. All candidates "took" the Numerical Ability test. Because reliability is based on the relationship between a candidate's true score and a candidate's estimated score, the true score and estimated scores were correlated. The square root of this correlation is the reliability of the test. The reliability for our Numerical Ability test was calculated to be 0.874.

Chapter 6: Country Adaptations and Comparisons

Localization Process

As noted previously, our Verify tests are designed to be used globally and translated into multiple languages, while ensuring equivalence across language versions. Localization is the process of culturally adapting and translating test content in such a manner that it measures the relevant construct (ability, skill, personality trait, attitude, etc.) equivalently in the source and target cultures, aiding in making appropriate personnel decisions and appears indigenous to the target culture. Initial localization includes series of qualitative steps designed to produce content in the target culture that is expected to achieve these goals. Localization confirmation is a series of statistical procedures designed to confirm that these goals have indeed been achieved.

The following briefly outlines both the steps in the initial localization process and the statistical procedures used in adapting the test for the target culture. Additional details about this process can be found in the document called: *SHL Assessment Localization: Best Practices and Practical Guidelines.*

Initial Localization

Content Decentering

Content developed in one culture will inevitably contain material that is specific to that culture. Examples of this type of material include references to specific geographic locations, personal names, organizational names, currency, and customs. Decentering is the process of removing or modifying culturally specific material in the source content which would not be readily familiar in other cultures (van de Vijver & Poortinga, 2005). Decentering is not the same as reviewing the material for its relevance to a particular target culture; rather it is a general preparation of the source content which makes localizing it into multiple target cultures more uniform and minimizes the number of cultural specific changes that will be needed. For example, if a question in the source material contained a reference to someone being on a baseball team, the reference could be changed to someone being on a sports team. Once content has been decentered, the process does not need to be repeated for each target culture localization.

Once initial decentering is done by a source culture test developer, the content is then reviewed by a test developer from another region to provide a second cultural perspective on the material. Any additional culture specific material is identified and modified or removed. The revised content is reviewed by a second U.S. test developer to confirm that the construct represented by each question has not been altered.

The decentering process helps address points C.1, D.1, D.4, and D.5 in the International Test Commission Guidelines for Translating and Adapting Tests (2010), which are summarized below.

Table 5. International Test Commission Guidelines for Translating and Adapting Tests

Context	
C.1	Effects of cultural differences which are not relevant or important to the main purposes of the study should be minimized to the extent possible.
C.2	The amount of overlap in the construct measured by the test or instrument in the populations of interest should be assessed.
Test Development and Adaptation	
D.1	Test developers/publishers should insure that the adaptation process takes full account of linguistic and cultural differences among the populations for whom adapted versions of the test or instrument are intended.
D.2	Test developers/publishers should provide evidence that the language use in the directions, rubrics, and items themselves as well as in the handbook are appropriate for all cultural and language populations for whom the test or instrument is intended.
D.3	Test developers/publishers should provide evidence that the choice of testing techniques, item formats, test conventions, and procedures are familiar to all intended populations.
D.4	Test developers/publishers should provide evidence that item content and stimulus materials are familiar to all intended populations.
D.5	Test developers/publishers should implement systematic judgmental evidence, both linguistic and psychological, to improve the accuracy of the adaptation process and compile evidence on the equivalence of all language versions.

Context	
D.6	Test developers/publishers should ensure that the data collection design permits the use of appropriate statistical techniques to establish item equivalence between the different language versions of the test or instrument.
D.7	Test developers/publishers should apply appropriate statistical techniques to (1) establish the equivalence of the different versions of the test or instrument, and (2) identify problematic components or aspects of the test or instrument which may be inadequate to one or more of the intended populations.
D.8	Test developers/publishers should provide information on the evaluation of validity in all target populations for whom the adapted versions are intended.
D.9	Test developers/publishers should provide statistical evidence of the equivalence of questions for all intended populations.
D.10	Non-equivalent questions between versions intended for different populations should not be used in preparing a common scale or in comparing these populations. However, they may be useful in enhancing content validity of scores reported for each population separately.
Administration	
A.1	Test developers and administrators should try to anticipate the types of problems that can be expected, and take appropriate actions to remedy these problems through the preparation of appropriate materials and instructions.
A.2	Test administrators should be sensitive to a number of factors related to the stimulus materials, administration procedures, and response modes that can moderate the validity of the inferences drawn from the scores.
A.3	Those aspects of the environment that influence the administration of a test or instrument should be made as similar as possible across populations of interest.
A.4	Test administration instructions should be in the source and target languages to minimize the influence of unwanted sources of variation across populations.
A.5	The test manual should specify all aspects of the administration that require scrutiny in a new cultural context.
A.6	The administrator should be unobtrusive and the administrator-examinee interaction should be minimized. Explicit rules that are described in the manual for administration should be followed.
Documentation/Score Interpretations	
I.1	When a test or instrument is adapted for use in another population, documentation of the changes should be provided, along with evidence of the equivalence.
I.2	Score differences among samples of populations administered the test or instrument should not be taken at face value. The researcher has the responsibility to substantiate the differences with other empirical evidence.
I.3	Comparisons across populations can only be made at the level of invariance that has been established for the scale on which scores are reported.
I.4	The test developer should provide specific information on the ways in which the socio-cultural and ecological contexts of the populations might affect performance, and should suggest procedures to account for these effects in the interpretation of results.

Target Cultural Adaptation and Review

The purpose of this step is two-fold. First, it is designed to confirm that the decentered material does not contain inappropriate, offensive or irrelevant material for a specific target culture. Although the decentering process should minimize the content not appropriate for a given culture, the uniqueness of cultures requires that the content be reviewed in light of each culture. This step addresses constructs, construct behaviors, customs, and conventions. For example, concerning constructs, this step may examine if “Coaching” is a relevant manager practice and, if so, is encouraging an employee in front of other employees a behavior that is reflective of good coaching. As necessary, surface features of the content were modified. An example of this would be changing a time format from a.m. /p.m. to a 24-hour format. Any modified questions were reviewed again by a source culture test development expert to confirm that the construct represented had not been altered. If a question could not be made culturally relevant by modifying surface characteristics, then the question was removed from the localized test.

Second, this step is designed to make surface modifications to the decentered material to make it more culturally acceptable to a specific culture. These modifications include issues such as inserting culturally specific personal names or local address/telephone number formats in graphics.

The target cultural adaptation and review process helps address points C.2, D.1, D.3, D.4, and D.5 in the International Test Commission Guidelines for Translating and Adapting Tests (2010).

Translation

Once the appropriate cultural adaptions are made, the initial translation of the source material into the target language is done. (In this context "translation" also refers to adapting text between dialects within a language, e.g., from U.S. English to U.K. English.) The translated test content material is then translated back into the source language by a separate translator. The back-translated material is then reviewed by source language test developers to confirm that the construct represented has not been altered. If it is judged that the meaning of any test content has been changed, the test developers confer with the translators to arrive at an appropriated modified translation. The final translation is then reviewed by native speakers for the target language to ensure the flow of the language.

For non-test content material (such as instructions and reports) and for within-language dialect translations, back translations are not conducted. Instead, after translation the material is reviewed by native speakers for the target language to ensure that language has a natural flow.

The translation steps helps address points D.1, D.2, and D.5 in the International Test Commission Guidelines for Translating and Adapting Tests (2010).

Localization Confirmation

Localization confirmation involves a series of statistical analyses that provide different types of evidence that the localized test is measuring the same construct across cultures and is functioning similarly in guiding personnel-related decisions. These analyses are conducted as sufficient data become available via content trialing or applied use.

Local Criteria-Related Validity

When sufficient test and job performance data are available a local validation study can be conducted. Job performance data may consist of supervisors' rating, objective performance measures or a combination of both. Establishing a significant and meaningful correlation between test scores and job performance measures in a sample from the target culture provides evidence that the test is predictive of job performance in the culture and is effective to use in selection decisions. This analysis helps address D.8 in the International Test Commission Guidelines for Translating and Adapting Tests (2010).

Measurement/Structural Invariance and Differential Item Functioning

When sufficient test data are available, various aspects of measurement equivalence of the test across cultures can be examined using multi-group confirmatory factor analysis procedures (Vandenberg & Lance, 2000). Aspects that could be examined include invariant covariance, configural invariance, metric invariance, scalar invariance, invariant uniquenesses, invariant factor variances, invariant factor covariance and equal factor means. Differential item functioning analyses can be used to examine if the test performs the same across the source and target cultures at the item level (Ellis, 1989). These types of analyses help address issues D.6, D.7, D.8, D.9 and D.10 in the International Test Commission Guidelines for Translating and Adapting Tests (2010).

Chapter 7: Criterion-Related Validity

Background

The criterion-related validation effort began in 2009 with the GCI Quantitative Ability and Verify Numerical Reasoning tests and continues as of the publication of this technical manual with the Verify Numerical Ability test. The validation project involved research partner organizations from a broad range of industries. The validation evidence presented here is a summary of all studies completed to date. Multiple criterion-related validity studies have examined the statistical relationships between numerical ability and job performance metrics. This accumulated validity evidence permits meta-analytic examination of the predictive nature of the test content. In this context, meta-analysis provides synthesized information about results of multiple studies that used the same or similar test content in a variety of settings to judge the overall value of implementing a test in a selection system. In this section, we describe our approach to meta-analysis and the validity results for our Numerical Ability test.

Universal Competency Framework (UCF) Mapping

It has been demonstrated via a wide body of literature that cognitive or reasoning ability is the most consistent predictor of job performance (Schmidt and Hunter, 1998). The Numerical Ability test is relevant to assessment where the following are critical aspects of the job or role (these requirements are taken from the SHL Universal Competency Framework or UCF and further details on the UCF and its validity are available in Bartram, 2005).

Table 6. UCF Mapping to Numerical Ability

Dimension Name UCF 25	Component Name UCF 93
Working with Numerical Information	Working with Numbers
	Using Mathematics

Meta-Analysis Method

Due to the importance of demonstrating criterion-related validity, our research scientists' work with client partners to conduct validation studies for all of our standard tests. All of these studies vary in terms of the job level studied, the industries covered, the tests included, and the criteria measured. Meta-analysis is a process of combining validity data from multiple studies into a single analysis (Hunter & Schmidt, 2004). Because most validation studies typically only include about 125 participants with both test data and criterion data, the validity estimates from a single study are susceptible to sampling error and the effects of statistical outliers. Meta-analysis combines the studies into one very large sample that reduces sampling error and lessens the impact of statistical outliers. Therefore, the validity estimates generated by the meta-analysis will more accurately represent true relationships in the general population.

We conduct ongoing validity studies with client partners. Though we are constantly completing validation studies, not every study is appropriate for inclusion in our meta-analysis. A validation study must include at least one test that is currently in use and at least one generalizable performance metric to be included in the meta-analysis. A job performance metric is considered generalizable if it was something meaningful outside of the specific client that provided the data. Many client organizations collect specific metrics that are of great importance to that organization, but do not generalize to other jobs within that job level or industry. Finally, some studies are excluded from the meta-analysis because of data quality issues. These issues include unacceptable reliability of criteria and lack of effort in supervisor ratings (e.g. everyone is rated the same in all categories or insufficient time is spent on the job performance rating survey). No study is ever excluded from the meta-analysis due to undesirable results.

Job performance metrics and assessment solution scores were obtained for participants in each validation study, and correlation coefficients were derived from the data. Correlations within each study were statistically corrected for criterion unreliability, as suggested by the original proponents and authors of the meta-analysis method in the field of Industrial/Organizational Psychology (Schmidt & Hunter, 1998). Conservative default criterion reliability estimates of .60 were used to make statistical corrections for unreliability in the supervisor ratings. This value is based on various sources, including average intraclass correlations (ICCs) across time intervals, typical client experiences with these types of criteria, and the Industrial/Organizational Psychology literature (e.g. Viswesvaran, Ones, & Schmidt, 1996). The input to the meta-analysis consisted of corrected correlation coefficients weighted by the sample size (*N*). The correlations included in the meta-analysis have not been corrected for range restriction. For details supporting the decision not to correct for range restriction, see the "Predictive versus Concurrent Studies" section below.

Additional analyses were conducted to determine the extent to which other variables may contribute to the meta-analysis results. The percent of variance attributable to sampling error was calculated to determine the extent to which unknown artifacts influence the predictor-criterion relationship. Experts (Hunter, Schmidt, & Jackson, 1982) suggested that if more than 75% of the variance can be attributable to statistical artifacts, then it can be reasonably concluded that results are generalizable rather than situation-specific. In addition, the 80% credibility interval was calculated to represent the range of correlations a client may expect for a given solution component. Note that for correlations based on only one study, the credibility interval is based on the study's sampling error as it is not possible to compute the percent of variance accounted for by sampling error or other moderators.

Though the meta-analysis contains hundreds of studies, not all of them will be included in the results presented in this technical manual. Meta-analytic correlation coefficients can be generated based on breakdowns of the data by job level, industry, or study type.

Each cell in the meta-analytic matrix represents an independent meta-analysis where the values for *k* (number of studies) and *N* (number of cases) refer to the subset of studies contributing to the meta-analysis for that pair of variables. For any one cell, the computations of the meta-analytic validity coefficient, variance due to sampling error, and the credibility or confidence interval are calculated using established methods of meta-analysis (Hunter & Schmidt, 2004). Estimates of these values for correlations of composites (e.g., overall scores) will, however, be more complex, as there often will not be single values of *k* and *N* that apply to all component measures in a composite.

Most meta-analyses in the literature focus on the relationship between one pair of constructs. In the validation of selection tools, this usually involves the pairing of one predictor construct and one type of criterion variable. Our use of meta-analytic techniques is not particularly unusual in that each cell in the full matrix is treated as an independent meta-analysis. For our purposes, however, it is crucial to generate meta-analytic estimates of the relationships among all of the variables. As mentioned above, this matrix is used to compute validity estimates for composites or overall scores based on component variables that may have been included in different combinations across samples.

Meta-Analysis of Validity Evidence for Numerical Ability

Global Cognitive Index – Quantitative Ability Validation Evidence

Identification of Validation Studies

Eighteen criterion-related validity studies across seven organizations conducted between 2008 and 2013 were included in the meta-analysis. Participants in these studies were from the following industries:

- Higher Education
- Telecommunications Services
- Retail
- Insurance

Job Performance Criterion Measures

All 18 validation studies conducted used a job performance rating survey (JPR) that included several different types of ratings. The job performance rating form included performance area ratings that aligned with work behavior dimensions of broad job performance and specifically to cognitive ability, and additional ratings of global performance that measured overall job performance.

The performance dimension rating items were presented on a 7-point scale (with an additional "cannot rate" response category. The global rating items had multiple-choice anchors appropriate to each item.

For ease of interpretation and comparison across studies, four criterion composites were computed. Exploratory factor analyses supported the factor structure of the composites. A job performance rating composite, referred to as a Performance Area Composite, was created by averaging all of the individual job performance area ratings. A second composite was created that included only the cognitively loaded job performance areas called a Cognitive Area Composite. The third composite called a Global Area Composite included the global job performance items. Finally, a composite of all items in the rating form was created called a Total Composite.

Predictive versus Concurrent Studies

We report validity coefficients separately for predictive and concurrent studies. In a concurrent study, predictor and criterion data are collected close in time, whereas in a predictive study, test scores are used for predicting future performance. In general, concurrent studies use existing job incumbents for providing both assessment data and performance data, whereas predictive studies use job applicants who take the assessment prior to hire, and at some later point in time, provide on-the-job performance data (Society for Industrial and Organizational Psychology, 2003). A predictive study in its purest form is one that follows the following procedure (Cascio & Aguinis, p. 146, 2011):

1. Measure candidates for the job.
2. Select candidates without using the results of the measurement procedure.
3. Obtain measurements of criterion performance at some later date.
4. Assess the strength of the relationship between the predictor and the criterion.

What we refer to in the tables in this technical manual as 'predictive studies' are not predictive in the purest sense. One primary deviation from the above procedure is that our predictive studies usually involve using results of the measurement procedure for making selection decisions (in contrast to what is suggested in step 2 of the procedure). In fact, most of our predictive studies use the results of the measurement procedure for selection decisions because we have existing concurrent validity evidence that supports using the tools for selection. However, by using the measurement procedure for decision-making, particularly in a top-down selection process, direct range restriction will occur in the predictor data, which has the well-known effect of reducing the observed validity coefficients (Cascio & Aguinis, 2011; Thorndike, 1949). The range in possible scores on the assessments for the predictive studies included in the meta-analysis should be smaller than the general population because only candidates with higher scores would be hired. This diminished variance reduces the correlations between a test score and a criterion. In this case, the most appropriate remedy is to correct the observed validity for range restriction, using the right correction given the way in which the scores were used when the restriction occurred (Sackett & Yang, 2000; Schmidt & Hunter, 1996). We do not correct our predictive or concurrent studies for either direct or indirect range restriction because clients have asked for clear guidance on what they can expect for an observed validity in a local validation study that utilizes the measurement procedure for selection decisions, without allowing for any type of correction.

In general, our predictive study results often are lower than that of concurrent studies because of uncorrected range restriction. The validity estimates presented for predictive studies in this manual are lower bound estimates, and will be conservative. Hunter, Schmidt, and Le (2006) estimate that for some predictors, the reduction in the true correlation due to range restriction could be as much as 25%. Even if range restriction was addressed, our concurrent designs are often likely to produce higher validity coefficients as this would be consistent with the broader research literature. Van Iddekinge and Ployhart (2008) summarize a number of studies showing that predictive designs usually tend to yield lower validities than concurrent designs, even those that have fully corrected for error due to unreliability and range restriction. They offer a few reasons for typically higher concurrent validities, including greater response distortion for applicants, thus distorting validities, the larger time lag between the collection of the predictor and the criterion data resulting in decrements in validity due to time, and the potential advantage that incumbents may have knowing more about what is expected in the job than inexperienced applicants, resulting in higher association between predictor and criterion. Although predictive validities are usually stronger in concurrent validation studies, this is not always the case.

In summary, we provide both estimates to give a range of validities to communicate to clients what they may expect for observed correlations if a concurrent validation study is conducted, or alternatively, a local validation study using a predictive design with selection.

Meta-Analysis Results

Tables 7 and 8 provide information on the meta-analytic validity of our Global Cognitive Index - Quantitative Ability test across all studies. The meta-analysis was conducted across different job levels: Entry Level, Professional / Individual Contributor, Supervisor, Frontline Manager, and Senior Manager / Director. Further, the overall meta-analytic statistics were generated for all of these studies combined. To help establish benchmarks for what size of correlation coefficients could be expected for different combinations of measures and outcomes in personnel research and practice, Bosco, Aguinis, Singh, Field and Pierce (2014) reviewed numerous studies that spanned a 30 year period. For knowledge, skills and abilities (including general cognitive ability) predicting job performance they found that correlations ranging from .13 to .31 could be considered "medium" with about 33% of observed correlations occurring in this range. For psychological characteristics (personality traits, emotional states, etc.) predicting job performance they found that correlations ranging from .10 to .23 could be considered "medium." Correlations below and above these ranges could be considered "low" and "high" respectively.

Overall, the test performed well in the prediction of job performance as rated by direct supervisors. The Quantitative Ability test predicted performance, global, and overall rating composites consistently with corrected correlations ranging from .03 to .29.

Table 7. Meta-Analytic Criterion-Related Validity Results for Quantitative Ability – Concurrent Studies

Criterion	Number of Studies (*k*)	Sample Size (*N*)	Observed Correlation (*r*)	Estimated Operational Validity (*ρ*)[1]	Percent of Variance Accounted for by Sampling Error	Credibility Interval Lower Bound	Credibility Interval Upper Bound
All Studies							
Performance Area Composite	18	2,430	.13	.17	91	.13	.22
Cognitive Ability Composite	13	1,458	.19	.24	71	.15	.34
Global Performance Composite	18	2,451	.09	.12	100	.12	.12
Total Composite	18	2,430	.12	.15	100	.15	.15
Entry Level							
Performance Area Composite	6	836	.17	.22	100	.22	.22
Cognitive Ability Composite	4	579	.18	.24	100	.24	.24
Global Performance Composite	6	831	.11	.15	100	.15	.15
Total Composite	6	830	.15	.19	100	.19	.19
Professional / Individual Contributor							
Performance Area Composite	7	1110	.12	.16	89	.12	.21
Cognitive Ability Composite	4	395	.23	.29	71	.19	.40
Global Performance Composite	7	1,137	.10	.13	100	.13	.13
Total Composite	7	1,117	.12	.15	100	.15	.15
Supervisors							
Performance Area Composite	2	172	.00	.00	66	-.13	.13
Cognitive Ability Composite	2	172	.07	.09	55	-.07	.25

Criterion	Number of Studies (*k*)	Sample Size (*N*)	Observed Correlation (*r*)	Estimated Operational Validity (*ρ*)[1]	Percent of Variance Accounted for by Sampling Error	Credibility Interval Lower Bound	Credibility Interval Upper Bound
Global Performance Composite	2	171	-.07	-.09	100	-.09	-.09
Total Composite	2	171	-.03	-.04	100	-.04	-.04
Frontline Managers							
Performance Area Composite	1	118	.23	.30	K = 1	.15	.44
Cognitive Ability Composite	1	118	.30	.38	K = 1	.24	.52
Global Performance Composite	1	118	.16	.20	K = 1	.05	.35
Total Composite	1	118	.21	.27	K = 1	.13	.42
Senior Manager / Director							
Performance Area Composite	3	290	.09	.12	45	-.06	.30
Cognitive Ability Composite	2	217	.14	.17	29	-.07	.42
Global Performance Composite	3	289	.02	.03	69	-.08	.14
Total Composite	3	289	.07	.09	51	-.07	.25

[1]Correlations in the Estimated Operational Validity column have been corrected for criterion unreliability.

Table 8. Meta-Analytic Criterion-Related Validity Results for Quantitative Ability – Predictive Studies

Criterion	Number of Studies (*k*)	Sample Size (*N*)	Observed Correlation (*r*)	Estimated Operational Validity (*ρ*)[1]	Percent of Variance Accounted for by Sampling Error	Credibility Interval Lower Bound	Credibility Interval Upper Bound
All Studies							
Performance Area Composite	5	639	.05	.07	100	.07	.07
Cognitive Ability Composite	1	103	.16	.21	K = 1	.05	.37
Global Performance Composite	5	690	.02	.03	100	.03	.03

Criterion	Number of Studies (k)	Sample Size (N)	Observed Correlation (r)	Estimated Operational Validity (ρ)[1]	Percent of Variance Accounted for by Sampling Error	Credibility Interval Lower Bound	Credibility Interval Upper Bound
Total Composite	3	430	.02	.03	100	.03	.03
Entry Level							
Performance Area Composite	2	252	-.01	-.01	100	-.01	-.01
Cognitive Ability Composite	1	103	.16	.21	K = 1	.05	.37
Global Performance Composite	2	252	-.05	-.07	100	-.05	-.05
Total Composite	2	252	-.03	-.04	100	-.03	-.03
Professional / Individual Contributor							
Performance Area Composite	1	178	.10	.13	K = 1	.01	.26
Global Performance Composite	1	178	.08	.10	K = 1	-.03	.22
Total Composite	1	178	.10	.13	K = 1	.00	.25
Frontline Manager							
Performance Area Composite	2	209	.08	.10	100	.10	.10
Global Performance Composite	2	260	.06	.07	100	.07	.07

[1]Correlations in the Estimated Operational Validity column have been corrected for criterion unreliability.

Verify – Numerical Reasoning Validation Evidence

Identification of Validation Studies

Six criterion-related validity studies across three organizations conducted between 2009 and 2013 were included in the meta-analysis. Participants in these studies were from the following industries:

- Banking
- Telecommunications Services
- Retail

Job Performance Criterion Measures

Five of the validation studies conducted used a job performance rating survey (JPR) that was mapped to the UCF dimensions. The JPRs also included additional ratings of global performance that measured overall job performance. The sixth study compared scores on predictor measures to an objective sales metric. These results will be presented separately in Table 9.

The performance dimension rating items were presented on a 5-point scale (with an additional "cannot rate" response category. The global rating items had multiple-choice anchors appropriate to each item.

Two criterion composites were computed. A job performance rating composite, referred to as a Performance Area Composite, was created by averaging all of the individual UCF performance area ratings. A second Global Area Composite that included the global job performance items was also created.

Meta-Analysis Results

Table 9 provides information on the meta-analytic validity of the Verify Numerical Reasoning test across all concurrent studies; no predictive studies have been conducted for this test to date. Overall, the test performed well in the prediction of job performance as rated by direct supervisors. The overall meta-analytic statistics were generated for all of these studies combined. The Numerical Reasoning test predicted performance and global rating composites consistently with corrected correlations of .17 and .18. The Numerical Reasoning test was also shown to predict an objective measure of call center performance in a telecommunications role, as seen in Table 10.

Table 9. Meta-Analytic Criterion-Related Validity Results for Numerical Reasoning – Concurrent Studies

Criterion	Number of Studies (*k*)	Sample Size (*N*)	Observed Correlation (*r*)	Estimated Operational Validity (ρ)[1]	Percent of Variance Accounted for by Sampling Error	Credibility Interval Lower Bound	Credibility Interval Upper Bound
Performance Area Composite	5	1,008	.14	.18	72	.11	.25
Global Performance Composite	5	1,007	.13	.17	35	.01	.32

[1]Correlations in the Estimated Operational Validity column have been corrected for criterion unreliability.

Table 10. Validation Results for Telecommunications Call Center Role

Criterion	Sample Size (*N*)	Observed Correlation (*r*)	Estimated Operational Validity (ρ)[1]
Total Sales/Number of Calls Taken	126	.11	.12

[1]Correlations in the Estimated Operational Validity column have been corrected for criterion unreliability using an estimate of .8

Chapter 8: Construct Validation

Overview

Construct validity provides information regarding the relationship of a test or assessment to other measures purported to measure similar or different constructs. Evidence of convergent validity exists when an assessment is highly correlated with another established measure of the same construct. Evidence of discriminant validity exists when an assessment has little or no correlation with another established measure of a construct that should have no theoretical link with the focal construct. Construct validity evidence is typically obtained by administering a number of similar and distinct measures in the same assessment battery, and then examining the relationships among the variables.

The construct validity of the Numerical Ability test has been examined in the context of its relationship with the GCI Quantitative Ability test, as well as the Verify Numerical Reasoning test. The validation study was executed by administering both tests (Verify Numerical Ability and GCI Quantitative Ability) together, to the same sample of examinees. The order of the tests was reversed halfway through data collection in order to balance the effect that test order may have had. Then the same process was repeated for the Managerial/Graduate and Supervisory/Operational versions of the Verify Numerical Reasoning tests. The correlation of the candidates' final theta estimates was then calculated. The results are presented in Table 11. There is no established standard for what should constitute a "good" correlation between two tests that purport to measure the same construct. The correlations below indicate that the pairs of tests share between 46.2% and 68.9% of their variance. The reason that the shared variance is not 100% is likely due to minor differences in how the construct is measured across the four tests. Also, Numerical Ability is adaptive and can cover a broader range of difficulty than either of the Numerical Reasoning tests.

To date no evidence of discriminant validity exists. Verify Numerical Ability has not been included in any projects that include assessments with which numeric ability should not be correlated. This chapter will be updated when such evidence becomes available.

Table 11. Construct Validation Correlation

Comparison	*N*	*r*
Bivariate correlation between Numerical Ability and GCI Quantitative Ability	865	0.69 (uncorrected) 0.83 (corrected)
Bivariate correlation between Numerical Ability and Numerical Reasoning M/G[1]	665	0.69 (uncorrected) 0.81 (corrected)
Bivariate correlation between Numerical Ability Numerical Reasoning S/O[2]	1,060	0.58 (uncorrected) 0.68 (corrected)

[1] M/G refers to the Managerial/Graduate form of the Numerical Reasoning Test

[2] S/O refers to the Supervisory/Operation form of the Numerical Reasoning Test

Chapter 9: Group Comparisons and Adverse Impact

Subgroup Differences

The purpose of the present analysis was to calculate subgroup difference effect sizes as a way of determining the potential for adverse impact towards females, racial/ethnic minorities, and candidates age 40 or older. Analyses were conducted on data collected on the Verify Numerical Ability test administered to candidates and employees in the United States in U.S. English and U.K. English versions of the instrument. In brief, the analyses indicate minimal to moderate differences between groups.

Sample

The sample consisted of 12,012 job candidates and employees who completed the instrument between July, 2013 and September, 2014 as part of a job application. As reporting demographic information is optional, only those who reported at least some demographic data were included in this analysis. Sample sizes used to calculate subgroup difference effect sizes are provided in Table 12. Effect sizes for subgroup comparisons were not reported when the subgroup sample size (*N*) was less than 200 because samples smaller than this are more susceptible to sampling error and typically lack the statistical power necessary to detect differences at critical *d* thresholds (Cohen, 1988**).** Therefore, results for the Hispanic and American Indian groups are not reported. Statistics for other racial/ethnic groups where data were not available will be updated upon collection of sufficient data.

Table 12. Sample Sizes Used for Effect Size Calculation

Group	Gender		Age		Racial/Ethnic Group				
Subgroup	Male	Female	<40 years	≥40 years	White	Black/ African American	Asian	American Indian/ Alaska Native	Hispanic or Latino
N	3,414	2,452	6,745	1,005	2,976	583	1,669	11	119

Analysis

Keeping with the Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice; 1978, Section 4D), between-group score differences and the potential for adverse impact (AI) were examined for the Verify Numerical Ability test. However, because cut scores will vary across organizations based on company needs, it is not possible to calculate the typical AI statistics of the 4/5^{ths} rule and the statistical difference in selection ratios (the likelihood that the difference in selection ratios is significantly greater than zero in standard error units - commonly referred to as the "2 standard deviation rule" or *Z* test), as referenced in the Federal Contract Compliance Manual issued by the Office of Federal Contract Compliance Programs (OFCCP, 2013). Instead, an analysis of the standardized mean differences (*d*) was conducted across different samples using a combined sample of candidates.

Statistically speaking, *d* values are more informative than the 4/5^{ths} or 2 standard deviation tests because they are pure effect sizes and not dependent upon pass rates or sample sizes. Across contexts, a *d* value of 0.2 is considered small, 0.5 is considered medium, and 0.8 is considered large (Cohen, 1988). Within the personnel selection domain, standardized mean differences of 1.0 are typical (or not uncommon) for Black-White differences on cognitive ability tests (see Sackett & Ellingson, 1997; Sackett, Schmitt, Ellingson, & Kabin, 2001). However, personality measures tend to exhibit small mean differences for race and gender (e.g., Hough, 1998; Ones & Anderson, 2002; Schmitt, Clause, & Pulakos, 1996).

Subgroup difference effect sizes were calculated on mean Verify Numerical Ability theta scores using the *d* statistic, the standardized mean score difference between groups. Negative *d* values indicate the protected or minority group scores below the referent group, and therefore may be a cause for adverse impact. An effect size of -0.5 or lower would have greater likelihood of producing adverse impact at most practical cut scores (Sackett & Ellingson, 1997, see Table 2, p. 712).

When interpreting standardized mean differences for the Verify Numerical Ability test, it should be noted that findings of adverse impact do not violate EEOC guidelines, provided the characteristic being measured is job relevant and no other tests are available that measure the same construct with less adverse impact. If the test demonstrates a predictive relationship with job-related criteria it is legally defensible (e.g., Uniform Guidelines on Employee Selection procedures; EEOC et al., 1978, Section 4D). With that said, other factors such as organizational goals, adverse impact for other possible selection assessments, and job characteristics should all be considered when determining test use. Table 13 contains the effect sizes for subgroup differences. A close examination of the table reveals that ds for subgroup comparisons are small to medium for age and gender, and large for race. Considering typical effect sizes of personnel selection tests, the Verify Numerical Ability test is somewhat likely to result in adverse impact for racial groups.

Table 13. Subgroup Difference Effect Sizes

Female*	≥40**	Black/ African American***	Asian ***
-0.21	-0.39	-0.98	0.10

*Referent group is Male.
**Referent group is <40 years old.
***Referent group is Caucasian.

Interpretation of effect size magnitude is given by Cohen (1988). Highlighted cells indicate effect size differences according to the scale shown in the table below:

Effect Sizes
Small to Medium: >|0.2| to ≤|0.5|
Medium to Large: >| 0.5| to ≤ |0.8|
Large: ≥ |0.8|

Adverse Impact and Cut Scores

In order to drive candidates to the next step in the selection process, clients typically utilize a cutoff score to differentiate between candidates who can be "Recommended" and "Not Recommended" at each stage in the process. While this cut score can vary based on selection process and candidate flow, most clients utilize a cutoff at the 25th or 30th percentile (i.e., screening out the bottom 25% or 30% of candidates based on our national norms or client-specific norms, when applicable). Our job solutions are configured by default to use the 30th percentile as a passing score. Based on our research and practical experience, this level will typically screen out the least qualified candidates from a candidate pool, reduce the potential for adverse impact, and provide good ROI for an organization in their screening program. Clients may refer to Sackett and Ellingson (1997, Table 2) and De Corte and Lievens (2005, Tables 1-3) to assist with interpreting effect sizes in relation to selection ratios and determine where to set selection cutoff scores to minimize adverse impact. Furthermore, we advise clients to conduct local adverse impact analyses whenever possible, as different testing situations, job levels, and candidate populations may alter the results from what is expected.

Conclusion

The subgroup difference analysis presented indicated small differences between gender and age groups, and large differences in racial groups, according to well-established professional guidelines for interpreting effect sizes. Specifically, Whites scored higher than Blacks, but the difference between Whites and Asian/Pacific Islanders was negligible. Males scored slightly higher than females. Those in the under 40 age group scored higher than those equal to or greater than 40. These results are consistent with established research indicating differences amongst race/ethnic groups when using cognitive ability instruments. Cognitive ability tests and tests with higher cognitive load will tend to show adverse impact (Sackett & Wilk, 1994). It is important to remember that findings of adverse impact do not violate EEOC guidelines, provided the characteristic being measured is job relevant, and no tests are available that measure the same construct with less adverse impact. Therefore, we make every effort to suggest valid tests that demonstrate the least adverse impact possible. Additionally, we recommend job analysis and local validation.

At this time the risk of adverse impact using the Verify Numerical Ability test appears moderate. However, although a particular test may demonstrate subgroup differences, it is more important if the selection process in its entirety demonstrates adverse impact. Because overall adverse impact can be reduced by the inclusion of other content that does not show adverse impact, we offer solutions with a broad base of valid content that balances optimal prediction with reduction in overall adverse impact. It must be noted that solution components can be weighted to essentially eliminate adverse impact, but this almost always reduces a solution's validity considerably.

Users are advised to monitor their use of the Verify Numerical Ability test for employee selection for indications of adverse impact, as stated by the Uniform Guidelines for Employee Selection Procedures (Equal Employment Opportunity Commission et al., 1978, Section 4D; Society for Industrial and Organizational Psychology, 2003).

References

Ackerman, P.L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*, 227-257.

Alfonso, V.C., Flanagan, D.P., & Radwan, S. (2003). The impact of the Cattell-Horn-Carroll theory on test development and interpretation of cognitive and academic abilities. In Flanagan, D.P. & Harrison, P.L. (Eds.), *Contemporary Intellectual Assessment, Second Edition: Theories, Tests, and Issues* (pp. 185-202).

Baddeley, A. (1986). *Working memory.* Oxford, UK: Clarendon Press.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2014). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*, 431 - 449.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Carretta, T.R. & Ree, M.J. (1996). Factor structure of the Air Force Officer Qualifying Test: Analysis and comparison. *Military Psychology, 8*, 29-42.

Carroll, J.B. (1993), *Human cognitive abilities: A survey of factor-analytic studies*, New York: Cambridge University Press.

Cascio, W. F., & Aguinis, H. (2011). Applied psychology in human resource management (7th Ed.). Upper Saddle River, NJ: Prentice Hall.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22.

Cattell, R., & Cattell, A. (1960). *Culture Fair Test*. Champaign, IL: Institute for Personality and Ability Testing.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

De Corte, W. & Lievens, F. (2005). The risk of adverse impact in selections based on a test with known effect size. *Educational and Psychological Measurement, 65,* 737-758.

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

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38290-29315.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74*,912–921.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

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

Galton, F. (1928). *Inquiries into human faculty and its development.* New York: Dutton. Reprinted from 1907 version; original work published 1883.

Graduate Management Admissions Council (2012). *The Official Guide for GMAT Review*. Hoboken, NJ.

Guilford, J.P. (1967) *The Nature of Human Intelligence.* New York: McGraw-Hill.

Horn, J.L. (1994). Theory of fluid and crystallized intelligence. In R.J. Sternberg (Ed.), *The encyclopedia of human intelligence* (Vol 1, pp. 443-451). New York: Macmillan.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57,* 253-270.

Hough, L. M. (1998). Personality at work: Issues and evidence. In M. Hakel (ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 131-159). Hillsdale, NJP: Erlbaum Associates.

Hough, L.M., Oswald, F.L., & Ployhart, R.L. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9,* 152–194.

Hunter, J.E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 257-266). Hillsdale, NJ: Erlbaum. Hunter, J.E. & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72-98.

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., (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed). Thousand Oaks, CA: Sage.

Hunter, J.E., Schmidt, F.L., & Jackson, G.B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, Ca: Sage Publications.

Hunter, J.E., Schmidt, F.L, & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*, 594–612.

International Test Commission (2010). International Test Commission Guidelines for Translating and Adapting Tests. [http://www.intestcom.org].

Johnson, M.F. & Weiss, D.J. (1980). Parallel forms reliability and measurement accuracy comparison of adaptive and conventional testing strategies. In D.J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Kanfer, R., Ackerman, P.L., Murtha, T., & Goff, M. (1995). Personality and intelligence in industrial organizational psychology. In D.H. Saklofske & M. Zeidner (Eds.), *International handbook of personality and intelligence* (pp. 507-622). New York: Plenum.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf–Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151−179). New York: Guilford.

McGrew, K.S. & Evans, J.J. (2004). Internal and external factorial extensions to the Cattell-Horn-Carroll (CHC) theory of cognitive abilities: a review of factor analytic research since Carroll's Seminal 1993 Treatise. *Carroll Human Cognitive Abilities (HCA) Project Research Report #2*. Evans Consulting: Institute of Applied Psychometrics.

Moreno, K. E., & Segall, D. O. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 169—174). Washington, DC: American Psychological Association.

Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology, 50*, 823-854.

Office of Federal Contract Compliance Programs. (2013). Federal contract compliance manual. Washington, DC: U.S. Department of Labor. Retrieved from http://www.dol.gov/ofccp/regs/compliance/fccm/FCCM_FINAL_508c.pdf

Ones, D. S., & Anderson, N. (2002). Gender and ethnic group differences on personality scales in selection: Some British data. *Journal of Occupational and Organizational Psychology, 75*, 255-276.

Raven, J. C., Court, J. H., & Raven, J. (1977). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. London: Lewis.

Ree, M.J. & Earles, J.A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1*, 86-89.

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. Personnel Psychology, 50, 707-721.

Sackett, P. R., & Yang H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*, 112-118.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education. American Psychologist, 56, 302-318.Sackett, P. R., & Wilk, S. L. (1994). Within group norming and other forms of score adjustment in pre-employment testing. *American Psychologist*, *49*, 929–954.

Sackett, P. R., & Wilk, S. L. (1994). Within group norming and other forms of score adjustment in pre-employment testing. *American Psychologist*, *49*, 929–954.

Salas, E., DeRouin, R.E., & Gade, P.A. (2007). The military's contribution to our science and practice: People, places, and findings. In L.L. Koppes (Ed.), *Historical perspectives in industrial and organizational psychology* (pp. 169-189). Mahwah, NJ: Erlbaum.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F. (2003). International Validity Generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology, 56*(3), 573-605.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199-223.

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.

Schmitt, N. & Chan, D. (1998). *Personnel selection: A theoretical approach.* Thousand Oaks, CA: Sage.

Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some job-relevant constructs. In c. R. Cooper & I. T. Robertson (Eds.). International review of industrial and organizational psychology (Vol. 11, pp. 115-140). New York: Wiley.

Segall, D.O. (1988). A procedure for scoring incomplete adaptive tests in high stakes testing. Unpublished manuscript. San Diego, CA: Navy Personnel Research and Development Center.

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

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Sociology, 15*, 201-293.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Stanovich, K.E. (1999). *Who is rational? Studies of individual differences in reasoning.* Hillsdale, NJ: Earlbaum.

Thorndike, E. L. (1949). Personnel selection: Test and measurement techniques. New York, NY: Wiley.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

United States Army Recruiting Command (2013). Frequently Asked Questions about the SIFT. Retrieved from http://www.usarec.army.mil/hq/warrant/index.shtml.

United States Department of Labor. (n.d.). The O*NET Content Model. Retrieved October 1, 2008, from http://www.onetcenter.org/content.html.

van de Vijver, F. J. R & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds*.), Adapting educational and psychological tests for cross-cultural assessment* (pp.39-63). Mahwah, NJ: Lawrence Erlbaum Associates.

Vandenberg, R.J., & Lance, C.E. (2000). A Review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3,* 4-70.

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology, 61,* 871-925.

Viswesvaran, C., Ones, D.S., & Schmidt, F.L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557-560.

Wagner, R.K. (1997). Intelligence, training, and employment. *American Psychologist, 52*, 1059-1069.

Yoakum, C.S. & Yerkes, R.M. (1920). *Army Mental Tests*. New York: Henry Holt & Co.

