Chapter 2:
Foundations for Measurement

John C. Scott
APTMetrics, Inc.

Alan D. Mead
Illinois Institute of Technology

Technology-enhanced assessment offers tremendous opportunities and unique challenges in the measurement and prediction of human behavior. By harnessing emerging technologies, organizations can reach across the boundaries of language and geography to accurately assess an almost limitless array of candidate attributes. Test users can now leverage sophisticated, web-based, assessment platforms to simulate any number of work environments and situations—effectively capturing candidates' ability to respond under real life conditions. These advances in technology have both demanded and facilitated the development of new measurement practices and theories (e.g., adaptive testing, item response theory) that have resulted in significant enhancements in assessment precision and efficiency. When used properly, automated assessments have the potential to provide a much more reliable, accurate and efficient means of measuring human characteristics than their erstwhile (paper-and-pencil) counterparts.

Despite the clear benefits and advances that technology-enhanced assessments bring to the table, there remain some key challenges that must be addressed to ensure alignment with sound measurement principles and practices. Increasingly, pressure has been mounting by a variety of test users to reexamine certain testing principles that they believe are limiting the full potential of automated assessments. One notable example relates to the use of un-proctored internet tests (UIT). Not so long ago, good testing practice would require a group of test takers (e.g., candidates for a job) to be assembled in a well-lit, distraction-free room with trained

proctors who would verify each test taker's identification, distribute the tests, read aloud the instructions, answer any questions, monitor the time limits, ensure test security and collect and log the tests and all associated materials. This standardized mode of administration was established to duplicate the procedures used in validating the test so that the results could be confidently interpreted for making sound decisions. For larger organizations that may test thousands or even tens-of- thousands of candidates a month, the logistics, time and costs associated with these sorts of standardized testing practices have led to questions regarding their real value and whether the rewards associated with violating a few established practices might in fact outweigh the risks.

There is no question that a clear business case can be made for the use of technology-enhanced assessments. In fact, as organizations begin to recognize the potential of automated assessments, their use will increase significantly and continue to expand on a global scale. The question then becomes how to achieve the right balance between a business's return-on-investment priorities with that of sound measurement practices so that critical assessment decisions can be made with efficiency, accuracy and integrity.

The purpose of this chapter is to address the measurement challenges – and highlight the opportunities – that technological advances bring to the assessment field. We begin by laying the foundation for sound measurement practice that will provide solid support for building and implementing high quality assessments. We then explore the importance of standardization and measurement equivalence in the context of automated assessments and reveal how cheating, response distortion and retesting can impact an assessment's psychometrics. We also address how computer access and the "technology divide" can impact performance on the assessment.

BUILDING HIGH QUALITY ASSESSMENTS

While the specific format of technology-enabled assessments can vary widely, there is a core set of underlying measurement principles that should be applied universally, regardless of how the assessment tools are configured and administered. Without the foundation of solid psychometrics to drive these assessments, the advantages that technology brings will ring hollow and the organization may actually be worse off than if it hadn't implemented an online assessment in the first place.

Since the focus of this chapter and this book is on the assessment of talent in organizations, we will direct our discussion to those measurement criteria required to successfully assess and predict behavior in the workplace. While organizations may decide to buy or build their assessment program, the measurement criteria described below apply to either decision.

The quality of any assessment can be evaluated by the extent to which it: 1) measures relevant criteria, 2) follows a clear set of assessment specifications, 3) provides a precise and consistent measure of the characteristics it is intended to measure, and 4) produces appropriate inferences (i.e., prediction) of behavior and performance.

Measure Relevant Criteria

The first step in developing (or purchasing) a high quality assessment tool is to clearly specify the constructs (e.g., knowledge, skill, ability, other personal characteristics; KSAOs) that need to be measured. This involves more than an informal review of job descriptions or anecdotal accounts of what it takes to be successful in a job. What is required, particularly when high-stakes testing (e.g., selection) is involved, is a well executed job analysis. Job analysis should serve as the foundation for any assessment program. Legal guidelines (*EEOC Uniform*

*Guideline*s, 1978) and professional standards and principles (*i.e.*, *APA Standards*, 1999; *SIOP Principles*, 2003) describe the importance of job analysis in the development of legally defensible, fair, and effective assessment programs.

There are a number of different approaches for conducting a job analysis that have evolved over the years and that are reflective of the dynamic nature of work and new organizational challenges. The choice of job analysis methods is driven by the purpose of the assessment (e.g., training diagnostic vs. hire/no hire decision), as well as practical and legal considerations, and there is no one preferred approach for all situations.

One way to determine how rigorous a job analysis should be to support a particular assessment application is to consider the level of risk involved should it be challenged. As the stakes increase, so does the level of rigor required in the job analysis. When assessment systems are challenged legally, the first area often investigated is the job analysis. At issue is how comprehensive and accurate the assessment criteria are to support the talent-related decisions. Unfortunately, many companies cannot produce solid job analysis data or documentation, and in many cases must conduct "post hoc" analyses when faced with a challenge to their assessment program. It is always most efficient and cost effective to conduct a robust job analysis as the first step in implementing any assessment program.

Develop Assessment Plan

Once the job analysis has been completed, the next step is to create an assessment plan that will clearly outline the attributes that need to be measured and identify the types of assessments appropriate for the targeted application. The assessment plan will establish the framework and specifications for determining: 1) the most appropriate item types and

administrative format, 2) how to properly construct the assessments and 3) how to ensure that the assessment results possess the required measurement properties.

As technology advances, the variation in testing formats becomes almost limitless. Emerging technologies that include interactive simulations, the use of avatars and virtual reality will all become readily available to creative test developers (Reynolds & Rupp, 2010). Computer adaptive testing (CAT), which is already well entrenched in larger testing programs, has made good use of both advancing technology and theory to provide highly reliable and innovative measures with far fewer items administered than would be required with traditional assessments. It is therefore important to account for the implications of these ongoing developments at the assessment planning stage, since it will impact the number and the nature of the items required. For example, while CAT administers fewer items, it actually requires a much larger pool of items than traditional testing formats to ensure adequate calibration across a range of ability levels.

The assessment plan should also account for user demands, such as the need to limit the length and administration time while also "engaging" candidates in the experience. These sorts of requirements have to be balanced with measurement considerations, such as the need to ensure adequate construct coverage and reliable results.

Build Assessment Specifications. The most effective way to ensure that an assessment is constructed to meet user demands while also accurately measuring the targeted attributes, is to develop a comprehensive set of assessment specifications. These specifications serve as a blueprint for the test developers and should draw upon the job analysis to systematically identify the topic areas to be assessed by the test and determine the relative weight that should be afforded to various KSAO areas within the assessment battery or single test. The specifications

should fully outline the content to be covered, the number of items to be included within each content area, the stimulus and response characteristics of the items (e.g., stimuli presented as pictures with associated audio – responses presented in a forced-choice format) and the administrative format. Table 1 shows an extract of how this component of the test specifications might be presented.

Insert Table 1 About Here

There are dozens of novel item types that have emerged over the past decade (Hambleton & Pitoniak, 2002; Zenisky & Sireci, 2002) and there is certainly no lack of creativity when it comes to leveraging technology to simulate tasks across a broad array of work environments. Theory about the targeted attribute should drive choices about the types of items that will best evoke examinees' demonstration of that attribute. For example, conceptualizing *Emotional Intelligence* as an ability would suggest using items that require the respondent to views photos or listen to recorded conversations and identify the emotions experienced by the actors. Those who conceptualize *Emotional Intelligence* as a personality trait, on the other hand, might use self-report or biodata items. The challenge for technology-enabled assessments will be to select from the wide range of potential stimulus and response options that are available to solicit a clear, job relevant and efficient demonstration of the targeted attribute.

Given the tremendous array of options afforded by technology-enhanced platforms, it is generally useful to have a guiding framework in mind when building specifications for innovative item types. Parshall, Davey, and Pashley (2000) developed an item taxonomy that can be helpful when organizing assessment specifications. They arranged item types along 5 dimensions of innovation: item format, response action, media inclusion, level of interactivity, and scoring algorithm. *Item format* refers to the type of response that is evoked from the

examinee. The two major types of item formats are selected response (e.g., multiple-choice) and constructed response (e.g., essay, video recording of answer). *Response action* refers to the mechanism used to provide responses (e.g., laptop camera, keyboard, touch screens). *Media inclusion* refers to whether and how video and audio are incorporated into the assessment. *Level of interactivity* refers to the extent to which an item interacts with or adapts to examinee responses (e.g., CAT vs. traditional) and the final dimension, *scoring algorithm*, refers to how the examinee responses are translated into score results. Parshall et al.'s (2000) taxonomy covers the key issues that need to be considered when blueprinting item types and formulating an assessment plan.

It is also important when building assessment specifications to include the expected distribution of psychometric indices (e.g., difficulty and discrimination levels) based upon the purpose of the test (e.g., mastery vs. selection). This is particularly important for CATs where the accuracy of ability estimates depends on a wide range of item difficulties within the item pool. The assessment specifications should also take into account whether or not the assessments will be proctored. If the test user plans on UIT, a large pool of items will be required so that they can be replaced on a regular basis and also used to populate any planned verification tests (Drasgow, Nye, & Tay, 2010; International Testing Commission, 2006). Finally, details about how the items will be scored should be clearly designated.

Incorporate Face Validity. One of the real advantages of technology-enhanced assessments is their ability to simulate key aspects of the work performed. Face validity is an important characteristic that should, whenever possible, be built into the assessment specifications. The reason that this is important is that, particularly in high stakes testing, examinees who believe that they are being assessed on characteristics relevant to the purpose of the test are more likely

to place credence on the measure and try their best (e.g., blueprint reading items on a selection test for an Architect job, customer service simulation items for a Customer Service job). Assessments that predict future job performance very well but don't look or feel like their intended purpose (interpretation of poetry passages for a technical job that requires reading comprehension) may give rise to a legal and/or labor relations challenge should the examinee perform below standard on the test. Including face validity is usually a fairly simple choice (one that we recommend to all test developers), and will be easier with technology-enhanced measures that can readily simulate realistic, work-related scenarios.

Conduct Editorial Review and Pretest the Items. Once the assessment items have been constructed, and before they are field tested, an editorial review should be conducted to ensure that the items are properly formulated (e.g., item stems are phrased as complete sentences, distracters "look" and "sound" like the correct answer). In the event that assessments are translated and will be used in other countries and cultures, it will be necessary to not only conduct a review of how well the assessment has been translated (see section on *Adaptation and Language Translation* later in this chapter), it is also recommended that an editorial board be convened that represents each country where the assessment will be implemented. This board should be tasked with ensuring that the actual intention or meaning of each of the items carries forward to the target culture. This review should be complemented by a field test that will provide a second level of analysis as to the fidelity of the translation.

Once the editorial review has been completed, the newly developed assessments should be field tested (this is independent of and as a precursor to a validation study) to ensure that the instructions are clear, the items are working as intended (difficulty and discrimination) and that the measures are reliable. The assessment plan should include the methodology for piloting these

items in the actual setting for which they will ultimately be administered (e.g., proctored small groups, un-proctored kiosks). A pilot is absolutely essential to evaluate the measurement properties on a representative tryout sample. This sample should include representation from groups protected by EEO laws (e.g., gender, race) within the U.S and multicultural/multilingual representation in the case of a global selection program. It will be particularly important that every examinee attempt every item so time limits are generous enough to minimize the number of "not-reached" items.

Table 2 provides an overview of the sorts of analyses that, at minimum, should be conducted on the pilot sample.

Insert Table 2 About Here

GATHERING PSYCHOMETRIC AND VALIDATION EVIDENCE

Once the assessment has been properly constructed and field tested, it is necessary to establish the psychometric and validity evidence needed to make accurate behavioral inferences. The challenge and the opportunity in the context of new assessment technologies, is to demonstrate that the measurement properties of novel item types and administrative formats justify their application. This section reviews classical and modern approaches for establishing reliability and provides recommendations for enhancing the precision of technology-driven measures. This section also discusses the strands of validity evidence that are needed to ensure adequate coverage of the targeted attributes and accurate prediction of job-related behaviors.

Establish Reliability

Traditional assessments create reliable scores by containing a large number of the best items (as shown by pilot testing) and then scoring each item independently. The challenge for simulations and other innovative item types, is the need to yield as many independent

measurement opportunities as possible. Reliability is typically lower for work samples and simulations since the time requirements of these items tend to limit the number of questions that can be administered. If a CPA candidate is to fill out a tax form that has 4 sub-schedules, and if a mistake in any sub-schedule will produce the wrong answer on the tax form, then provision must be made for partial credit or else the entire tax form simulation will be essentially one "item" on a traditional assessment. Scoring algorithms must also allow for normal variations (e.g., deductions can be summed in box 11 or itemized in boxes 11a through 11e). Because measurement opportunities often cannot be easily dropped (like independent multiple-choice items), psychometrically poor items are often kept but weighted zero in the scoring. While simulations and work samples tend to provide more measurement information than traditional multiple-choice tests, they offer less information per minute of testing time than multiple-choice items (Jodoin, 2003), and therefore their results will generally be less reliable for time-limited administrations.

The basic requirements of reliability transcend administrative format and apply to all forms of assessments where the objective is to produce an accurate measure of the targeted attribute. The key question in the context of this chapter is whether and how much technological innovations and associated practices impact the scope and magnitude of measurement error. For example, one might argue that the increased administrative flexibility afforded by UIT would most certainly increase measurement error, but that might be offset by the precision of a set of items presented in an adaptive format. The critical issue here is determining the major sources of error, estimating their size and ideally, identifying strategies that can leverage the technology to improve reliability. As the stakes and consequences of assessment decisions increases, so does the importance of reliability.

There are two psychometric theories in use today that drive our assumptions and approaches for estimating reliability: random sampling theory and item response theory (Bejar, 1983). Random sampling theory – which continues to be popular and in wide use – includes both classical testing and generalizability theories. This theory defines measurement error as the extent to which an individual's observed scores on an assessment randomly deviate from his/her hypothetical true score. The objective here is to determine how well an observed score generalizes to the universe from which it is drawn, and approximates the true score. The larger the measurement error, the less confidence we have in generalizing beyond the observed scores and specific test.

It should be noted that reliability and standard error of measurement (SEM) estimates that are calculated through these random sampling theory procedures only apply to the test scores and not the assessment itself. That is, reliability is considered an attribute of the test data and not the assessment, so it is inappropriate to ever state that the assessment itself is reliable. In fact, the APA *Standards* (1999) state that when reliability is reported, it must be accompanied by a description of the methods used to calculate the coefficient, the nature of the sample used in the calculations and the conditions under which the data were collected. All of these caveats are necessary due to the fact that the reliability estimates calculated through these procedures are sample dependent, and as a result, have a number of practical limitations when building, or evaluating, technology-enhanced assessments.

Use Item Response Theory (IRT) to Replace Single Index of Reliability. It is necessary in high stakes testing to be able to determine how well a test discriminates along the ability continuum, particularly around the critical values used to set the cutoff scores (APA *Standards*, 1999). IRT allows us to calculate measurement error to this level of precision by replacing the

concept of reliability with that of the *test information function.* The test information function tells us how precisely each ability level is being measured by the test. One of the challenges in using IRT, and what has prevented more widespread application of this theory, is the sample size requirement for calibrating item parameters. For example, for a 60 item test, a sample size of 1,000 is generally required for stable parameter estimates using the 3-parameter model. This is generally not a problem for large testing programs but may be so for those applications that have only a few hundred cases. Fortunately, most technology-enhanced assessments are deployed because of high-volume hiring so the use of IRT becomes more feasible.

Under the IRT conceptualization, the relationship between ability (θ) and the probability of success on an item P_i (θ) can be expressed in the form of an Item Response Function (IRF) as shown with 5 separate items in Figure 1. The probability of passing the item falls on the vertical axis, and the ability continuum (i.e., the "theta scale") falls on the horizontal axis. As ability increases, so does the probability of passing the item.

Insert Figure 1 About Here

This relationship between ability (θ) and the probability of success on an item P_i (θ) can also be expressed as:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp\{a_i(\theta - b_i)\}} \quad (1)$$

The *a* parameter is the item discrimination index and represents the steepness of the IRF. The *b* parameter, which represents item difficulty, is defined as the point on an ability scale at which the probability of a correct response to an item is .5. The *b* parameter has the same metric (is on the same scale) as θ so the difficulty of an item can be directly compared to the ability of a test-taker. Item 1 in Figure 1 is the easiest and farthest to the left on the theta scale while item 5

is the most difficult and the farthest right. The *c* parameter indicates the probability that an examinee with very low ability will get the item correct and is often called the guessing parameter. It functions as the lower (or left hand) asymptote of the IRF.

In terms of estimating an examinee's ability, θ, not all items are equally effective. IRT provides the *item information function*, $I_i(\theta)$, to show how effective an item is at measuring a given range of ability. The definition of item information is quite technical (the squared rate of change in the probability of a correct response divided by the variance of the item, as shown below). However, the use of information functions is quite simple.

$$I_i(\theta) = \frac{\left[P_i'\right]^2}{\sigma_i^2} \tag{2}$$

Figure 2 shows the item information functions for the items plotted in Figure 1.

Insert Figure 2 About Here

Notice that where the IRF's rise steeply, information is high, while information is low where IRF's are flat—indicating that the item is ineffective at measuring examinees in that range of theta. Each item has a maximum degree of information and a range on the theta scale where it is effective. Item information functions sum to create the test information function, $I(\theta)$:

$$I(\theta) = \sum I_i(\theta) \tag{3}$$

Test developers should construct tests to have high information over the important ranges of the theta scale (or over the entire theta scale) by selecting those items yielding the most information. When IRT is used in this way, test length can be minimized without sacrificing

measurement precision (especially using adaptive testing, described at the end of this section). In Figure 2, the bold line shows the test information (the sum of the individual item information functions). Although real tests would have more items, Figure 2 illustrates how tests can be constructed to have uniformly high information over the entire range of scores: The items must have a good spread of item difficulties and each item should have good item discrimination.

The degree of precision of the IRT test score, $\hat{\theta}$, can be calculated from the test information function. [Theta-hat, $\hat{\theta}$, is an estimate of the person parameter, θ, and is the IRT "score" for a test-taker.] The conditional standard error of measurement of $\hat{\theta}$ is the square root inverse of the test information function:

$$SE(\hat{\theta} \mid \theta) = \frac{1}{\sqrt{I(\theta)}} \tag{4}$$

Figure 2 also shows the relationship between the test information curve and standard error of measurement for a 5 item test. The SEM is the "U-shaped" dashed line. The SEM curve is obviously a mirror image of the test information function. This means that imprecision/error of measurement is greater for scores at the edges of the score scale and is at a minimum across most of the score scale. Most tests have comparatively peaked test information functions because item difficulties tend to cluster around the center of the score scale. Good, general-purpose tests will look as close to Figure 2 as possible.

For pass/fail tests that have a known cut-score, the optimal assessment will have a test information function that peaks over the cut-score and may be quite low for other scores. This recognizes that on pass/fail assessments, only scores that determine whether a person passes or fails are important. On a driver's licensing exam, for example, only the score that determines

passing or failing is important to measure precisely; it is not helpful for that test to distinguish good from excellent (because both groups pass) or poor from very poor drivers (because both groups fail). Therefore, if we plan to use a single cutoff score in a selection context, a shorter test can be built by selecting only those items that are most informative at that specific ability level.

Computer adaptive testing combines advances in computer technology and IRT to create a very narrow, highly psychometric kind of artificial intelligence that can efficiently deduce the ability level of examinees from their responses with far fewer items than a traditional test. In fact, with a large pool of items calibrated using IRT, substantially shorter tests can produce more reliable scores. As testing is increasingly computerized and as item response theory becomes widely-used, many assessment programs will encounter fewer barriers to its use and realize significant incremental benefits from adaptive testing.

It should be noted that many selection tests and non-cognitive assessments (e.g., personality and attitude measures) have complex factorial structures and require multidimensional IRT models for item calibration. Multidimensional (MCAT; Segall, 1996) or bi-factor (BFCAT; Weiss & Gibbons, 2007) models provide a better basis for adaptively administering these assessments. Multidimensional models allow adaptive tests to leverage the correlation among traits—when someone responds in an introverted manner, they are slightly more likely to be conscientious as well. For example, in one simulation study of the adaptive administration of the 16PF Questionnaire, a unidimensional CAT allowed a reduction of test length of about 25% with only slight loss of reliability (Mead, Segall, Williams, & Levine, 1999). However, test length on the MCAT could be reduced to about 50% with similar, small loss of reliability.

Bi-factor analysis is used for constructs with a main general factor and specific indicator factors (for example, general intelligence or personality measures like 16PF Extraversion, which is thought to be composed of Interpersonal Warmth, Liveliness, Social Boldness, Forthrightness, and Group Affiliation). In one application of BFCAT, Weiss and Gibbons (2007) examined a 615-item personality instrument that had an overall score and four content scores. On average, the BFCAT reduced test length by about 80% with slight loss of reliability.

Establish Validity

The fact that technology-enhanced assessments can be created to so closely simulate activities performed on the job sometimes raises questions by organizational stakeholders as to whether there is really a need to formally validate the tool. Since the assessment "obviously" measures elements of the job and validation studies can be a time consuming and costly activity, what is the purpose of holding up implementation and delaying the dividends that the system could be paying? It is therefore not unusual to see validation placed low on the list of priorities by impatient stakeholders who may consider this activity more of a formality. However, despite the organizational pressures and advances in technology and measurement theory, the requirements for validation have not changed. The method may vary based upon the nature and purpose of the assessment (McPhail & Stelly (2010), but validation is never optional.

According to the APA *Standards* (1999) "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed use of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests "(p. 9). In the talent selection context, the intended use of an assessment is to predict job performance. Therefore, we are interested in two facets of validity: 1) how well the assessment measures the criteria that underlie successful job performance and 2) how well the assessment actually

predicts job performance. Guion (1998) refers to the first facet as *Psychometric Validity* (which subsumes content and construct validity) and to the second as *Job Relatedness* (i.e., criterion-related validity).

Evaluate Psychometric Validity. In the case of simulations and work samples, the most frequently applied and generally most practical approach for gathering evidence of psychometric validity is through a content validity study. The objective here is to evaluate the extent to which the KSAOs measured by the assessment represent the targeted content domain, which is determined through the job analysis and fleshed out through the test specifications. A measure of a test's content validity is generally not statistical (although expert ratings may be collected), but rather determined through agreement by subject matter experts that the items used are representative of the domain from which they were sampled. The necessary ingredients for building evidence of content validity include a comprehensive job analysis, thorough test specifications, competent test construction and expert agreement that the test content is related to and representative of the content domain.

Gather Evidence Based on Internal Structure. For more traditional personality and multiple-choice cognitive ability tests, the interrelationships between items on the test – often assessed through factor analysis – can be an effective way to determine how well the structure of the assessment matches the intended framework. A review of the item statistics (item difficulty and discrimination) will also help determine if the structure of the assessment supports the intended use. High item total correlations and internal consistency measures (e.g., coefficient alpha) provide evidence that the test scores are systematically measuring some variable. If the content is based upon a well structured job analysis and the test has internal consistency, it is reasonable

to assume that the items are measuring the intended attribute without contamination (Guion, 1998).

An analysis of *differential item functioning* (DIF; Holland & Wainer, 1993) is another means for determining whether the items that comprise the assessment are operating as intended and support the assessment's internal structure. By reviewing DIF across different subgroups (e.g., English- and Spanish- speaking shift supervisors for a multi-national organization) with similar ability – or standing on an attribute – differentially functioning items can be identified for follow-up review and modification as necessary.

Evaluate Job Relatedness. The most direct way to evaluate how accurately an assessment can predict important job-related criteria is to conduct a criterion-related validation study. Evidence of job-relatedness is determined through a correlation between the assessment and the criteria of interest. Other forms of evidence can also be leveraged to support job relatedness under certain circumstances (see McPhail, 2007 for a detailed description of alternative validation strategies including transportability, validity generalization and synthetic validity).

The choice of the performance criterion measures is of central importance in the validation study (APA *Standards,* 1999), and they must be held to the same psychometric validity standards used to evaluate the assessment measures (Guion, 1998). As is the case with assessment measures, the criterion measures must be based upon a comprehensive job analysis and appropriately reflective of the multidimensional nature of job performance. Flaws in selection decisions can occur through too narrow a conception of the facets of job performance that contribute to success – and subsequently – missed opportunities to account for these facets by the assessment tools. (Outtz, 2010).

Assessments should be validated under the actual conditions for which they will

ultimately be administered. For example, if the intent is to administer the assessment in an un-proctored setting, the validation study should be set up to mirror these conditions. Likewise, if a verification test will be implemented to confirm the results on the un-proctored test, and it will be administered under proctored conditions, the validation study of this test should occur in a proctored setting.

When a criterion-related validation study is properly conducted, the resulting evidence allows us to make informed decisions around how to maximize the prediction of performance, where to set passing scores to balance the goals of utility and fairness and how to implement a legally defensible selection program. The key criteria when evaluating criterion-related validity evidence are: 1) coverage of the important job performance criteria, 2) psychometric quality of test and criterion measures, and 3) relationship between predictor(s) and criteria (McPhail & Stelly, 2010).

STANDARDIZATION AND EQUIVALENCE

In a vault in the basement of the International Bureau of Weights and Measures on the outskirts of Paris, there sits a small cylinder of platinum and iridium—the *International Prototype Kilogram*, which has defined the meaning of "one kilogram" since it was manufactured in 1889. Copies of this standard exist in government bureaus around the world to enforce a standardization that allows a businesswoman in Beijing to know that the 20 kilograms of gold being offered by a dealer in London are equivalent to 20 kilos being offered in New York. Such standardization is essential for commerce and scientific progress in the physical sciences.

Standardization is also extremely important to psychological measures, where the construct being measured is unobservable and has no natural metric. If a personality test is being

used to select workers, it is critical that it produce the same measurements on Monday and on Friday, this year and next, and when administered on computer or on paper. It is also critical, in many instances, that it produce interchangeable scores when administered to English-speaking Canadians and German-speaking Swiss and all the other languages used in locations where a multinational organization recruits professionals, managers, and salespeople. *Equivalence* is the degree to which standardization is maintained when an assessment is changed. Thus, the topics of this section, standardization and equivalence, are very important foundation topics for technological assessment.

Standardization Characteristics

High-quality assessments are standardized—they ask each respondent to react to the same set of carefully-chosen questions or tasks under prescribed conditions designed to minimize irrelevant influences (e.g., quiet rooms, adequate lighting, comfortable environment). Administration in a noisy, uncomfortable place might lower scores due to these distractions and *not* due to real differences in the knowledge or ability. Similarly, if a military assessment designed to measure performance under pressure (using loud recorded sounds, violent role-players, etc.) were administered without such distractions, scores might well be significantly higher, *but not because the examinees were more tolerant of stress*.

Some researchers (e.g., Weiss, 2007) have criticized web-based testing because standardization can be much harder, or impossible, with this media. However, some evidence (Buchanan, Johnson, & Goldberg, 2005; Stanton, & Rogelberg, 2001) suggests that merely administering a test on a website does not preclude psychometric validity. If computerization makes reading the items harder, or changes any other influential characteristic of the examination process, then the computerization itself may affect standardization. Because the characteristics

that influence the examination process are not well understood, assessing equivalence is an important process.

Showing Equivalence

High-quality technology-enabled assessments are characterized by their equivalence across different conditions and groups. Measurement equivalence is related to standardization in that poor standardization, or violations of administration procedures, can produce non-equivalence. There are many other potential causes of non-equivalence. For example, research described below suggests that merely computerizing most kinds of assessments does *not* automatically cause non-equivalence. However, a bad interface design or very restrictive computer platform (e.g., a hand-held computer with a 4cm display or a tiny thumb keyboard) might introduce factors to the test that are irrelevant to the intended content. It would be inappropriate to compare people tested using paper-and-pencil to those tested with computerized tests that were not equivalent.

Relevant standards require test developers and users to show equivalence of paper and computerized forms of assessment. The APA *Standards* (1999) require that equivalence evidence be collected. The ITC *Computer-Based and Internet Delivered Testing Guidelines* (2005) are even more detailed, requiring that test developers show that computerized and paper forms have comparable reliabilities, correlate with each other at the level expected based on the reliability estimates, correlate comparably with other tests and external criteria, and produce comparable means and standard deviations or have been appropriately calibrated to render comparable scores (p. 11).

There are two main paradigms for researching equivalence: multiple-groups and multiple-measures and, as described above, there are three critical areas of equivalence. First, the

computerized and paper forms should rank order test-takers similarly. This requirement ensures that computerized and paper forms have similar reliability and measure the same construct, and can be shown statistically by correlating the scores of the computerized and paper forms of the test. Second, the mean and variance of test scores should be similar (either because of perfect raw-score equivalence, or because form-specific norms are used). Finally, scores from computerized and paper forms of a test should have similar correlations with important external criteria, such as job performance.

If the mean or variability of scores on the computerized form are different from those of the paper form, then separate norms or *equating* (Kolen & Brennan, 2004) can be used to create interchangeable scores if the two forms have construct equivalence. Usually, the adjustments are fairly simple, such as adding or subtracting a few points.

Multiple-groups equivalence designs. In the multiple-groups paradigm, one group takes one assessment (e.g., computerized) and another group takes the other assessment (e.g., paper). If the groups are randomly assigned, then any important ("statistically significant") difference in the mean scores for the two groups is taken as an indication of non-equivalence. The spread of scores for the two groups might also be compared, to see if one of the groups has a wider range of scores.

One serious problem arises if the groups are *not* randomly assigned to take one or the other form. If the groups are not randomly equivalent, then this design is seriously compromised because differences in test scores may well be due to group differences rather than with the form of the test taken. For example, if an attitude survey was administered on paper to day shift employees and on computer to night shift employees, what portion of the results are due to differences in the attitudes of day- and night-shift personnel? It is impossible to tell.

A technical problem with the multiple-groups equivalence designs is that hypothesis testing was designed to detect differences and is ill-suited to detecting equivalence (i.e., standard hypothesis testing cannot be used to support the Null hypothesis of no difference). Misusing hypothesis testing in this way has a number of unfortunate outcomes and should be avoided. While Rogers, Howard, and Vessey (1993) describe a framework for testing a hypothesis of equivalence, it would be best in these circumstances is to discard hypothesis testing and rely on effect sizes (or equating).

A more fundamental problem with the multiple-groups approach is that we cannot correlate the scores on the two forms (e.g., we do not have any information about whether people who scored well on the paper version also scored well on the computerized version). The correlation of scores on the two forms is *the* central issue in any equivalence research because it directly measures the degree to which the two forms of the assessment are reliably measuring the same thing. The multiple-groups paradigm is unable to address this question. Even worse, one could find that two quite different, and highly non-equivalent, assessments happen to have similar means and that two highly-equivalent assessments happen to have different means. Thus, this design detects only one kind of non-equivalence and the kind of non-equivalence is the kind easily handled by separate norms or equating.

Thus, we are skeptical about equivalence research that uses a simple experimental approach and basic null hypothesis significance testing to compare paper and computerized groups. An alternative approach is to test for measurement equivalence (MEQ) using structural equation models (SEM; Ployhart, Weekley, Holtz, & Kemp, 2002; Meade, Michels, & Lautenschlager, 2007) or item response theory differential item functioning (IRT DIF; Raju, Laffitte, & Byrne, 2002). This approach can be used to test whether relationships between items

and external criteria are the same across groups. Although the MEQ approach also cannot correlate scores across forms, it tests whether the items of the computerized and paper forms have identical psychometric properties (i.e., the same difficulty and pattern of correlations with other items). Using this approach, it is assumed that if the paper and computerized forms are measuring different things, then the item psychometrics would not be exactly the same across the forms. The *Adaptation and Language Translation Issues for World-Wide Assessment* section below describes the SEM MEQ and IRT DIF approaches.

Multiple-measurements equivalence designs. The alternative paradigm is the multiple-measure design, so-called because each volunteer is assessed with each of the forms (i.e., measured two or more times). For example, all examinees might complete both the paper and computerized versions of a scale (a single group takes both forms of the assessment). Although this design has important methodological advantages (e.g., allows the researcher to correlate the scores on the two forms), there are unique problems that may arise through this design. The main issue is the influence of the repeated testing. It is best to administer parallel forms on different days, counter-balancing the order of administration.

What level of correlation shows equivalence? If the "true scores" of the test-takers are the same (to within a linear transformation) on the computerized and paper forms, then the observed correlation will be attenuated by the reliabilities of the forms. Equation 5 shows how estimated reliabilities can be used to estimate the true-score correlation of the computerized and paper forms (the so-called "disattenutated" correlation).

$$r(X,Y) = r(T_X, T_Y)\sqrt{r_{XX}\, r_{YY}} \tag{5}$$

In this equation, *r (X, Y)* represents the "observed" correlation between the scores on the

predictor (i.e., test) *X* and the criterion, *Y*, $r(T_X, T_Y)$ represents the correlation between true scores and r_{XX} and r_{YY} are the reliabilities of X and Y. Because reliabilities are values less than one, the observed validity is always less than the true-score validity (the observed validity is said to be "attenuated by measurement error in X and Y").

If the estimated true-score correlation, $r(T_X,T_Y)$, is 1.0 then the construct being measured by the two forms is perfectly equivalent. [Note that even if the correlation is 1.0, the forms may have different means or variences and equating may be needed; however, if the equivalence correlation is low then no analysis can possibly produce equivalent forms.] Values below 1.0 indicate lower degrees of equivalence and, because they are correlations, are usually easily understood by psychologists and other test users. [Values above 1.0 should not occur; however, estimated correlations can exceed 1.0 due to sampling error. Byrne (1998) discusses this "boundary parameter" issue.] For example, values below .707 indicate that less than half of the variability in the scores on the paper and computerized forms are shared across the formats (see Mead & Drasgow, 1993).

Equivalence of Cognitive Assessments

Cognitive assessments have correct and incorrect answers and measure knowledge, skills, or abilities. In one of the earliest empirical comparisons of computerized and paper forms of an exam, military researchers (Sacher & Fletcher, 1978) administered vocabulary and logic tests to recruits in both computerized and paper formats. Their design allowed for the calculation of both the reliabilities of (both forms of) the test scores and the correlation of the scores across computerized and paper formats. The true-score correlation for 115 recruits was 0.95 and 0.87 for the vocabulary and logic tests respectively. Although these researchers found other issues (e.g., differences in response latency and answer changing), these correlations show excellent

comparability for the vocabulary test and good comparability for the logic test.

The logic test required recruits to answer six items per minute and so was considered fairly speeded, which likely impacted the comparability findings. Greaud and Green (1986) published an early and influential study of computerizing a speeded test, and they found poor equivalence. Thus, from the earliest research on this topic, speededness of the test emerged as a moderator of the comparability of computerized and paper forms.

The findings that speeded tests were less comparable should not be surprising because similar effects have been seen when seemingly small changes are made in the way that responses are recorded for speeded paper-and-pencil tests. For example, Boyle (1980) compared four groups who were all taking paper tests but using different kinds of optical marking answer sheets. He found that answer sheet formats requiring a single stroke were significantly different from a format that required a rather larger circle to be filled in—presumably a single stroke is a substantially different response then darkening a relatively large circle.

As more equivalence studies appeared in the literature, review articles also appeared to summarize the findings. In their influential narrative review of the literature, Mazzeo and Harvey (1988) suggested several possible moderators of equivalence, some of which have been subsequently discredited (e.g., ability to change answers) and some which have been supported (e.g., speededness). Bugbee (1996) provided another early narrative review that raised concerns about a lack of equivalence across media of administration in educational settings. However, a more recent narrative review by Paek (2005) concludes that K-12 students have access to computers in the classroom, frequently use computers for learning activities, and are comfortable with current technology. She concludes that the preponderance of the evidence supports equivalence except where long reading passages are present.

Mead and Drasgow (1993) published the first meta-analytic review and probably the most positive. For "timed power" forms (i.e., forms that were not highly speeded), they found a disattenuated correlation between paper and computerized formats of 0.97, which they interpreted as showing considerable support for the equivalence of carefully-developed computerized versions of cognitive ability tests which were not highly speeded (e.g., the GRE). When they examined highly speeded tests, they found a disattenuated correlation of only 0.72, which is a high correlation but clearly different from 1.0. (About half of the variance in the true scores of examinees were due to the computerization!) Also, equivalence of highly speeded tests was far more variable than for power tests. Mead and Drasgow interpreted this as support for speededness as a moderator of equivalence. Thus, it is particularly important to assess the equivalence of paper and computerized versions of highly speeded tests.

A few studies of computerization of speeded tests that have been published since the Mead and Drasgow (1993) meta-analysis have shown very good comparability between paper and computerized versions. Neuman and Baydoun (1998) showed essentially perfect true-score correlations for ten speeded clerical selection tests. Pomplun, Frey, and Becker (2002) studied computerized and paper versions of a speeded reading test and found a true-score correlation of 0.94. It is not clear whether computerized speeded tests are becoming more comparable to their paper counterparts (perhaps because of greater care taken by test developers, changes in the types of speeded tests studied, or because of advances in technology) or because of a file-drawer bias in published results, or some other reason. However, in a recent, carefully-designed comparison of web- and paper-based speeded forms (Mead, 2010), we found a cross-mode true-score correlation of 0.80—close to the 0.72 value found by Mead and Drasgow (1993).

Equivalence of Non-Cognitive Assessments.

Researchers have also examined the comparability of paper and computerized versions of non-cognitive predictors, such as attitudes, personality, measures of motivation, and automated interviews (i.e., intake interviews). Early comparability concerns focused on changes in socially-desirable responding, omitting items (Biskin, & Kolotkin, 1977), or anxiety (Canoune & Leyhe, 1985) caused by medium of administration. Of course, almost everything about computers and our relationship to computerization is different today, as compared to the 1970's and 1980's when computers were comparatively primitive and uncommon.

In one large meta-analytic investigation of socially-desirable responding on computerized, non-cognitive ability measures, Richman and her colleagues (Richman, Kiesler, Weisband, & Drasgow, 1999) examined differences between computerized and traditional formats across 61 studies and 693 means. Overall, they found an overall effect size of 0.02, which is very small, meaning that computerization matters very little.

Other recent, large-scale, analyses have suggested fairly good comparability. Meade and his colleagues (Meade, Michels, & Lautenschlager, 2007) used a structural equations measurement equivalence framework to compare Occupational Personality Questionnaire (OPQ) personality scales in a large sample of undergraduates. Their results generally suggested that the scales of the OPQ functioned equivalently when administered on paper or on the Internet. Curiously, however, they found better equivalence when participants could choose the medium of administration than when they were assigned to a medium.

Mead and Blitz (2003) reported a meta-analysis of multiple-measures studies of comparability. They found 105 studies comparing paper- and computer-based versions of non-cognitive assessments, mainly attitude or personality scales. However, only *six* studies used the multiple-measures design. Across 41 correlations from these studies, in a sample of N=760, they

found an overall true-score correlation of 1.02, which they interpreted as strong evidence for the comparability of non-cognitive abilities across administration modes.

Summary of Equivalence Issues and Recommendations. Research on assessments of both cognitive ability and non-cognitive constructs suggests that carefully developed computerized versions can measure the same construct as their paper counterparts—except for assessments with extensive reading or that are highly speeded, which may be noticeably less comparable. These results are good news because computerized tests that successfully measure the same construct should have the same criterion predictive relationships shown for paper forms. However, a final question remains—are separate norms needed for the computerized and paper forms?

Mead and Drasgow (1993) meta-analyzed the standardized mean differences between paper and computerized forms of timed power tests. They found an overall mean of -0.03, indicating that computerized power tests were very slightly more difficult than their paper counterparts. However, the estimated sampling error of these differences was 0.15, indicating that one could easily sample a mean difference of 0.15, 0.20, or even 0.30 for a given assessment. Similar results were obtained by Richman and her colleagues (Richman, et al., 1999) in an analysis of the socially desirable responding on non-cognitive measures. Thus, we recommend that unless research has shown that a given computerization did not affect the norms of a paper form, the computerized form have its own norms.

Adaptation and Language Translation Issues for World-Wide Assessment

For large multinational employers, technology-enhanced assessment enables the global use of assessment systems on an unprecedented basis. Given the far-reaching talent consequences brought about by these technological advances, the need to properly adapt the

assessment to the new context (e.g., a new country, region, culture) cannot be overstated. Modifications to the assessment across contexts may range from minor issues (introducing UK English spelling and metric units) through the removal of cultural idioms to the translation of the assessment and instructions into a new language.

Some programs use translation/back-translation (TBT) to detect translation quality. Because TBT has not been systematically studied, its effect on translation quality is not empirically known, although its wide use (despite substantial cost) may suggest that TBT does have some value. However, there are several reasons to be skeptical that TBT is sufficient. First, the translators must be bilingual and thus likely to have had substantial experience with the other culture. Second, some terms may translate poorly, yet in a way that back-translates well. Finally, the content of the instrument may interact with the culture of the respondents (Ryan, Horvath, Ployhart, Schmitt, & Slade, 2000; Liu, Borg, & Spector, 2004). A job satisfaction item that asks about one's boss might be perceived quite differently in high vs. low power-distance cultures. Questions about co-workers in collectivist cultures may be affected by in- and out-group issues that matter little to individualistic respondents. So, we strongly recommend that measures are pilot tested in the original and target culture(s) and measurement equivalence analyses conducted to detect such issues.

Measurement Equivalence for Adaptations. There are two widely-used frameworks for assessing the measurement equivalence of adapted tests, structural equations modeling (SEM) and IRT differential item functioning (DIF). Vandenberg and Lance (2000) provided an early review that clarifies how the SEM approach to measurement equivalence is far more nuanced (i.e., complex) as compared to the IRT DIF approach, which focuses very closely on the equivalence of the item difficulty and psychometric quality across adapted instruments (for

detailed comparisons, see Raju et al., 2002 and Stark, Chernyshenko, & Drasgow, 2006). The SEM approach is best when the response variables are fairly continuous (i.e., item responses should be on 5- or 7-point Likert scales) and multivariate normal. Otherwise, an IRT DIF approach may be better. SEM may also be preferred because it can simultaneously assess equivalence across multiple groups (most IRT DIF approaches only analyze two groups, so they have to be used in pairwise comparisons of multiple groups).

The SEM approach is quite flexible and is easily extended to assess the equivalence of item means (this model is sometimes called *Mean And Covariance Structures,* or MACS; see Ployhart & Oswald, 2004). An interesting limitation of MACS analysis is that the item difficulties and group means cannot simultaneously be assessed. Analysis of changes in the item means requires that the analyst assume that the groups have equal means and analysis of differences in the group means requires that the analyst assume that the items have equal difficulty across groups. IRT DIF approaches have a clever solution—if *most* of the items function equivalently, then IRT DIF approaches can separate the effects of a few item's difficulties changing from group differences. Empirical comparisons of the SEM/MACS and IRT DIF approaches suggest that they often reach similar conclusions when carefully similar analyses are conducted (see Stark et al., 2006) but there are also many instances of divergence due to different sensitivities of various IRT DIF statistics (see Raju et al., 2002).

The IRT DIF approach works well for categorical data, especially dichotomous responses from ability tests. IRT models are fit independently to each group and then a step called *iterative item linking* (Candell & Drasgow, 1988) is used to make the independent scalings comparable (incomparable scaling is very much like temperature measured in Celsius and Fahrenheit—the same construct but the temperatures cannot be compared until one converts to a common scale).

Various IRT DIF methods can then be used to evaluate the comparable item scalings; see Raju and Ellis (2002) for a practical review of several IRT DIF approaches.

CHEATING, RESPONSE DISTORTION AND RETESTING

In this section, we focus on *the effect* of cheating, response distortion and retesting on the psychometric properties of technology-enabled assessments. For a full discussion of cheating and response distortion, see Chapter 4 in this Volume by Arthur and Glaze. *Cheating* on an assessment refers to any deliberate, malfeasant means of altering one's assessment score—that is, any attempt to obtain a higher score by improper, deceptive, or fraudulent means. *Response distortion* refers to cheating (most typically by inflating scores) on a self-report measure—for example, a person responding "Strongly Agree" on a personality survey item asking "I never miss deadlines" when, in fact, missing deadlines is a common occurrence for that person.

In theory, cheating and response distortion might completely destroy the validity of assessment scores. If all candidates completing an assessment obtained scores different from their natural score, the correlation of these scores with a criterion would probably be attenuated, perhaps to zero. Interestingly, because the reliability of assessment scores is affected by random error and because cheating and response distortion might decrease random error, the reliability could actually seem to improve. However, this artifact simply shows that the validity of assessment scores is more important than the reliability of those scores.

Cheating and response distortion threaten validity in at least two ways. First, they may result in most people scoring about the same (technically, true-score variance is being diminished). When everyone scores very similarly, it is much harder to determine who are the best candidates—imagine a horse race where horses' noses are all within a few millimeters of each other; it would be difficult to determine the winner, even by photograph. We want

assessments that allow individuals to express their natural differences in an area; cheating and response distortion act against this and diminish the value of assessment scores.

Also, if some people are cheating and others are not, the cheaters will tend to rise to the top of rankings of the candidates. Of particular concern are those low-ability candidates who obtain a high score by fraudulent means and rise dramatically to the top scoring band. If the top scoring candidates are selected, then they could disproportionately be cheaters (Zickar, Rosse, & Levin, 1996). If the assessment scores are (otherwise) valid, then that suggests that these cheaters will have poor outcomes (low tenure, poor performance, etc.). When these low potential performers are selected along with candidates who obtained legitimate high scores, and who are therefore high potential, the selected group's mean job performance will be lower and assessment scores will be less useful and have lower operational validity than it might otherwise have been.

So, how bad is cheating and response distortion? Does it completely invalidate an assessment and automatically and invariably reduce the value of the assessment to zero? Not necessarily. In fact, validity coefficients are surprisingly robust to these issues (Ones & Viswesvaran, 1998; Rosse, Stecher, Miller, & Levin, 1998) because they take into account all of the scores from all of the candidates of an assessment. A small proportion of individuals obtaining higher assessment scores than they should (i.e., resulting in lower job performance than their scores would otherwise indicate) does not necessarily produce lower validity coefficients, particularly when the overall impact of cheating or response distortion is diluted over a large pool of candidates. Also, if all candidates distort their responses in the same way, the validity coefficient will be unchanged unless score changes cause “ceiling” or “floor” effects (where too many candidates get the best or the worst score), in which case the practical

usefulness of the assessment might be severely compromised.

Clearly, applicant response distortion can seriously affect the norms, and practitioners should ensure that they use norms that were collected under conditions similar to the context in which the assessment is to be deployed. Applicant norms should always be preferred, especially for non-cognitive measures such as biodata and personality instruments, where response distortion is common under high stakes conditions.

The effect of cheating on the psychometric properties of assessments is difficult to quantify, as it is ultimately dependent upon the proportion of test-takers out of the total pool that actually cheated. This number in turn is dependent upon the level of exam security, the degree of organization among the cheaters, the difficulty of the exam, the controls put in place to minimize cheating, and the degree to which the outcome impacts candidates' lives. It is axiomatic that cheating reduces the usefulness of an assessment in predicting job performance. The consequences to an organization of even a single poor hire can be substantial when you consider the costs associated with training, low productivity, turnover and the ultimate need to replace this individual. Multiply this by even a small number of low-ability cheaters hired into the organization and it becomes readily apparent that every effort should be taken to minimize opportunities to cheat on high-stakes exams. In particular, when high stakes assessments are administered under unproctored conditions, it is highly recommended that the results be verified under secure, proctored conditions. The verification version of the assessment should be validated under proctored conditions.

Retesting

Retesting on the same form is widely assumed to be detrimental to exam security. For example, if an unqualified candidate knows that he/she will be re-tested with the same form,

he/she may use the first assessment opportunity to memorize difficult items, which they will solve (or get their friends to solve) at home and memorize so that they achieve an inappropriately high score. This becomes particularly problematic in unproctored testing situations where candidates may be allowed to take the test an unlimited number of times under assumed names before they actually submit their responses for scoring. Besides alternate forms, counter-measures include controlling exposure of items through item inventory control mechanisms and monitoring re-test scores so that unusual score increases (e.g., more than two standard errors of measurement) can be investigated (see Chapter 4).

In one provocative study, researchers arranged for actual returning candidates to randomly receive either the same or a different form of a radiography examination (Raymond, Neustel, & Anderson, 2008). They observed virtually the same score increase of about half a standard deviation (d = 0.52 for those who received the same form versus d = 0.48 for those who received an alternative form). They did notice a small difference for administration time; those who received the same form took slightly less time (d = -0.02) while those who received an alternate form took slightly longer (d = 0.20). The authors note that the examinees had no reason to expect to be re-tested with the same form—if same-form retesting were to become common, one could imagine greater exploitation by candidates. Also, the context of this study might be unique for two reasons: First, the exam was very easy to pass (slightly over 90% passed on the first try). And, second, the nature of most of the exam items involved scrutinizing medical images. Thus candidates may not recognize items that they failed and opportunities to memorize items and study them are fairly limited with this item type.

In a meta-analysis of many such studies of aptitude and achievement tests, Hausknecht and his colleagues (Hausknecht, Halpert, Di Paolo, Moriarty, & Gerrard, 2007) found that repeat

examinees generally re-tested better but those who re-tested using the same form improved twice as much as candidates completing an alternate form (d = 0.45 vs. d = 0.24), although this difference decreased as the time between testing and re-testing lengthened.

Very little is known about the effect of re-testing on the validity of an assessment. Lievens, Buyse, and Sackett (2005) examined medical studies admissions tests and found that validity was higher for those re-taking a knowledge test and passing on the second attempt than those passing on the first attempt. However, validity was lower for an intelligence test. The authors suggest that knowledge can be studied and so higher performance on the repeat administration of the knowledge test reflects greater learning but higher scores on the intelligence test repeat administration simply reflects good luck, test-taking skills, etc. that are unrelated to the criterion. More research, especially on the prediction of actual job behavior, is needed in this area. Candidates who re-take an assessment can be expected to improve their scores. If the exam content can be "studied" then re-testing may produce larger score changes and an identical form may inflate this effect.

FAIRNESS OF TECHNOLOGY-ENHANCED ASSESSMENTS

There is no doubt that web-based assessments expand the reach of organizations to access a larger, more diverse, candidate pool (Beaty, Grauer, & Davis, 2006). However, legitimate concerns have been raised that not everyone has the same access to, or comfort with, computer technology and this may impact assessment outcomes and introduce fairness concerns – along with measurement error (Tippins, Beaty, Drasgow, Gibson, Pearlman, & Segall, 2006). Since the application of technology-enhanced assessments will undoubtedly continue to expand at an exponential pace, it is prudent to examine the potential impact that this delivery option has on groups with limited access or familiarity with this sort of technology.

Internet Access and the Digital Divide

The "digital divide" is a term that is used to describe the gap between those individuals who have access to various telecommunications technologies and those who do not. The component of that technology that is most applicable to online assessments is high-speed internet access or *broadband*. A recent survey found that 63% of Americans have broadband at home and that broadband availability is available to more than 90% of households (Horrigan, 2009). This study also found that senior citizens and low-income Americans had the largest gains in broadband subscriptions between 2008 and 2009, while African Americans experienced a below average broadband adoption growth rate in 2009, totaling 46% of current households.

Another study found that internet usage rates (defined as occasional usage) vary by race: 71% of Whites, 60% of Blacks, and 56% of Hispanics. (Fox & Livingston, 2007). The rate for Spanish-dominant Hispanics drops to 32%. This study also examined the effects of education on internet usage and found that internet use is uniformly low for those who have not completed high school: whites (32%), Hispanics (31%), and African Americans (25%) and uniformly high (about 90%) for those who have completed college (Fox & Livingston, 2007). Internet usage rate also declines with age: 91% for 18-30 year olds, 90% for 31-42 year olds, 79% for 43-61 year olds, 56% for 62-71 year olds, and 29% for those 71 and older (Rainie, Estabrook, & Witt, 2007).

While it is obvious that Internet access and usage is becoming more widespread, it is still not universal and there do appear to be some race and age differences. This raises potential ethical and fairness concerns that the use of assessment technology could result in differential subgroup performance (Pearlman, 2009; Tippins, et al., 2006). It is therefore incumbent on test users to review their situation and ensure that all candidates are treated fairly and afforded an

equal opportunity to demonstrate their standing on the attributes being assessed. The demographic data on broadband access and use can be informative when analyzed against the targeted pool of applicants.

From the standpoint of the assessment, there are a number of ways to enhance familiarity with the technology and ultimately elicit the best possible performance from the candidate. For example, a tutorial can be incorporated that will show prospective candidates how to navigate through the screens and respond to the different types of test items. This tutorial should incorporate sample items for each of the content areas being measured. An online narrator or "testing assistant" can also be programmed into the assessment to respond to common Q&As and even read the questions and responses if desired. In addition, help desk information should be made available to the candidates should they run into technical difficulties as they progress through the assessment.

For individuals who do not have access to, are uncomfortable with, or need special accommodations for online assessments, organizations should be prepared to offer supervised sessions whereby candidates can receive verbal instructions and assistance with technical issues. In the case of accommodation, it may be necessary in some circumstances to offer alternatives to the online assessment depending on the nature of the impairment (Tippins, et al., 2006). Additionally, an equated paper-and-pencil version of the test could be made available if there is a large enough segment of the applicant pool that could benefit from this alternative.

While there is no doubt that as access and familiarity with broadband increases over time the impact of technology as a moderator of assessment performance will steadily dissipate. However, for the time being, organizations should analyze how the use of UIT or other technology-enhanced applications are impacting their candidate pools and target appropriate

recruiting efforts to address any emerging gaps with the relevant labor market (Tippins, et al., 2006).

CONCLUSION

The application of new technologies to the field of assessment has resulted in a tremendous amount of innovative practice and leading-edge research. Test users are able to leverage this assessment technology and supporting research to implement tools that are scalable on a global level and which can measure a broader array of attributes and behavior. Candidates can be assessed in multiple languages, in remote locations, for any level of job, and this can be accomplished with fewer items and greater precision than ever before. Technology has helped reignite the popularity of assessment and through its efficiencies and wide-scale application, can produce returns on investment that are hard for organizational leaders to ignore.

These advantages and large-scale applications have also led to unique challenges associated with established measurement practices. Some of these challenges, such as the use of UIT or the desire to shorten assessments, have resulted in innovative solutions such as CAT and the application of sophisticated theories such as IRT. However, there remain some basic measurement tenets that need to be applied universally, regardless of the content or technological medium within which the assessments are administered. In the rush to beat the competition in the war for talent, technology-enabled assessment systems may sometimes be “stood up” without the necessary attention to these measurement principles.

Whether the assessment system is purchased or developed from scratch, it still needs to be able to reliably measure the targeted attribute(s) and make accurate inferences about work behaviors. Any blueprint for building and implementing high quality assessments must include at its core a thorough job analysis, detailed assessment plan and well designed field research to

establish the necessary psychometric and validation evidence. In addition, technology-enabled assessments must be implemented in a manner that ensures some level of standardization or the results may be suspect and of diminished value. Measurement error is tied to those irrelevant influences that interact with the test taker, so every effort should be taken to establish and follow prescribed administration procedures.

Cheating, response distortion, retesting and differential access to technology all impact the measurement accuracy of an automated assessment. It is therefore incumbent upon the test user to determine what level of impact each of these elements has in his/her own environment and to take appropriate action to address these sources of measurement error.

Good testing practice transcends the medium or application and the best way to fully leverage emerging technology is to ensure such assessments have a solid measurement foundation.

Table 1

Extract from Practical Reasoning Test Specifications

Candidates will be presented with modules consisting of multiple pieces of information. This information will be presented in different formats (e.g., tables, charts, graphs, text, etc.) and appear to come from different sources (e.g., memos, newspapers, books, manuals, etc.). Candidates will be required to:

1. answer specific questions about the details contained in the material;
2. evaluate the consequences associated with the information presented;
3. sift through the information to identify what is critical for taking action or making a decision;
4. take action based on the information; and
5. interpret and use the information to solve practical problems or situations.

Stimulus & Response Attributes

The information will focus on practical, business-related issues drawn from critical incidents provided by Subject Matter Experts. During the online, multimedia test, candidates take on the role of a first-line supervisor and are presented with a variety of "real life," on-the-job situations. These situations take place in areas including an operations center, a plant control room, a work site, and customer call center. Candidates must determine how they would respond to work situations based on information presented through live-action scenarios and interactive information resources. This test will include full-motion video where candidates are provided with access to an entire desktop as though they are sitting at their desk. Interruptions (e.g., phone calls, voice messages) will be built into the process as they would be on the job.

Candidates will be required to comprehend, evaluate and apply the information to solve problems, make decisions, and/or take action. There will be a total of 30 items in this section. Each item will have four response alternatives. Each alternative will plausibly relate to the content of the item stem. The correct answer will be based on accurate interpretation of the materials presented. Distracter response alternatives for items will be based on inappropriate or inaccurate interpretation of the information.

Skills Assessed

Performance on this test will be determined by candidates' ability to:

1. extract relevant information from tables, charts, graphs, and text to solve practical problems;
2. access, evaluate, and utilize information contained in manuals or other reference materials to make decisions, answer questions, or provide input to others;
3. synthesize information from various sources and communicate relevant information to others;
4. understand and apply new information, procedures, or principles to perform the task at hand; and
5. attend to and verify the accuracy and completeness of detailed information in documents or on the computer.

Table 2

Pilot Test Analyses

- P-Values (Item Difficulty - Proportion who responded correctly to test items)
    - Ensure item p-values match overall goal of test (e.g., mastery test would contain majority of items with relatively high p-values (.8); selection test would have average p-values in the .5 range)
    - Test items for which everyone responds correctly or incorrectly do not help us distinguish between test takers
    - Test items in which two or more response alternatives have high p-values may indicate the presence of more than one "correct answer"
    - Low p-value for the "correct" answer may indicate the correct answer is ambiguous or, in fact, incorrect
    - Test items that have low p-values for "correct" answers or that have two or more high p-values should be resubmitted to subject matter experts for their review

- Biserial or Point Biserial (Item Discrimination)
    - The Point Biserial (item-total score correlation) will be higher (closer to 1.00) when high scoring examinees get the item right and low scoring examinees get the item wrong
    - Ensure items possess good discriminating power (differentiate between high and low performers)
        - Positive, high-item total correlations are desirable

- Distracter Analysis
    - Ensure test items possess only one correct answer

- Review Overall Test Statistics
    - Mean
        - Describes overall difficulty (or easiness) of test
    - Standard Deviation
        - Describes distribution of test taker's test scores
    - Reliability
        - Reliability should exceed .80
    - Standard Error of Measurement (SEM)
        - SEM should be considered for use in "banding" around the cut-off to take into account test's measurement error

Figure 1

Relationship between ability (θ) and the probability of success on an item P_i (θ)

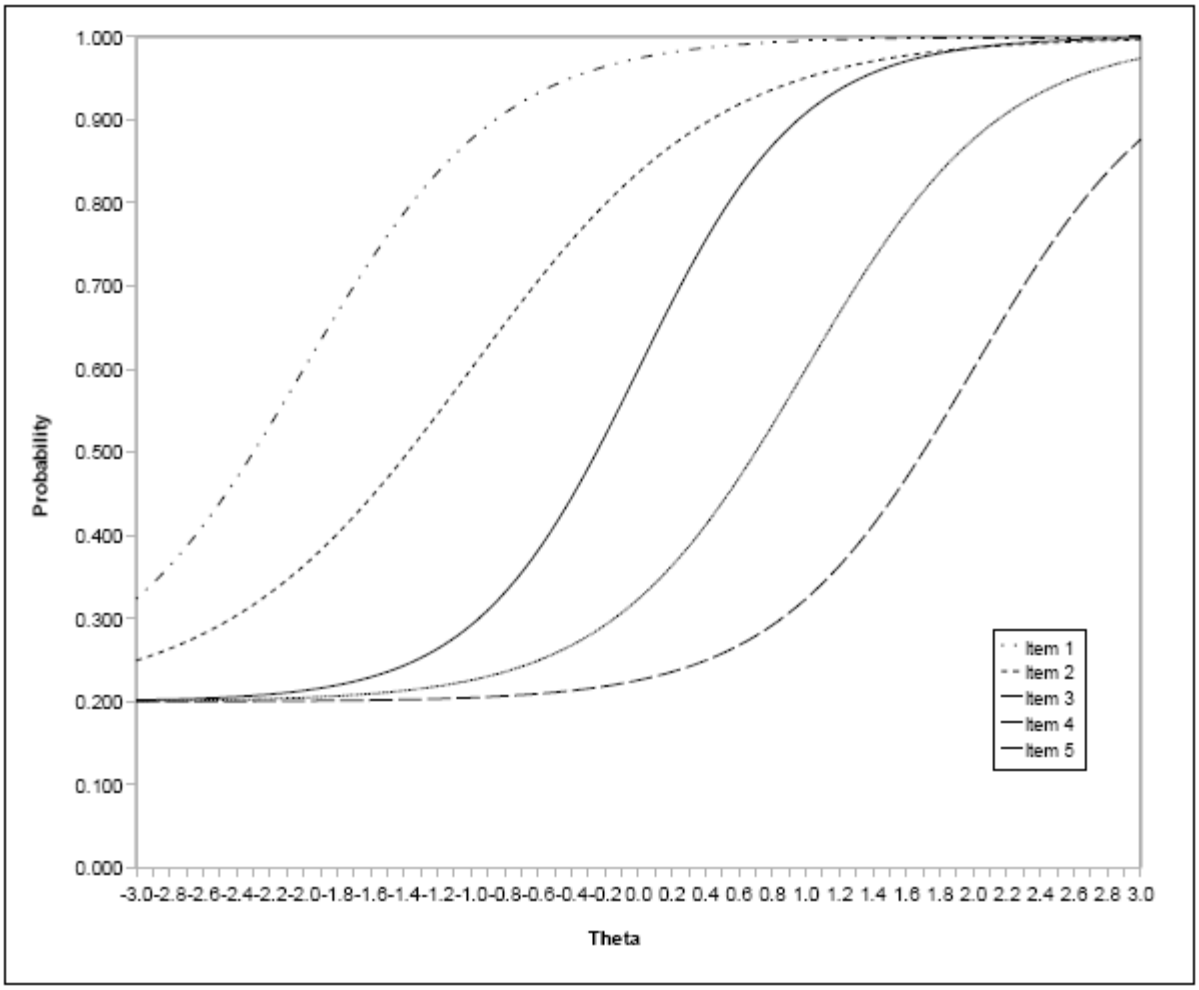
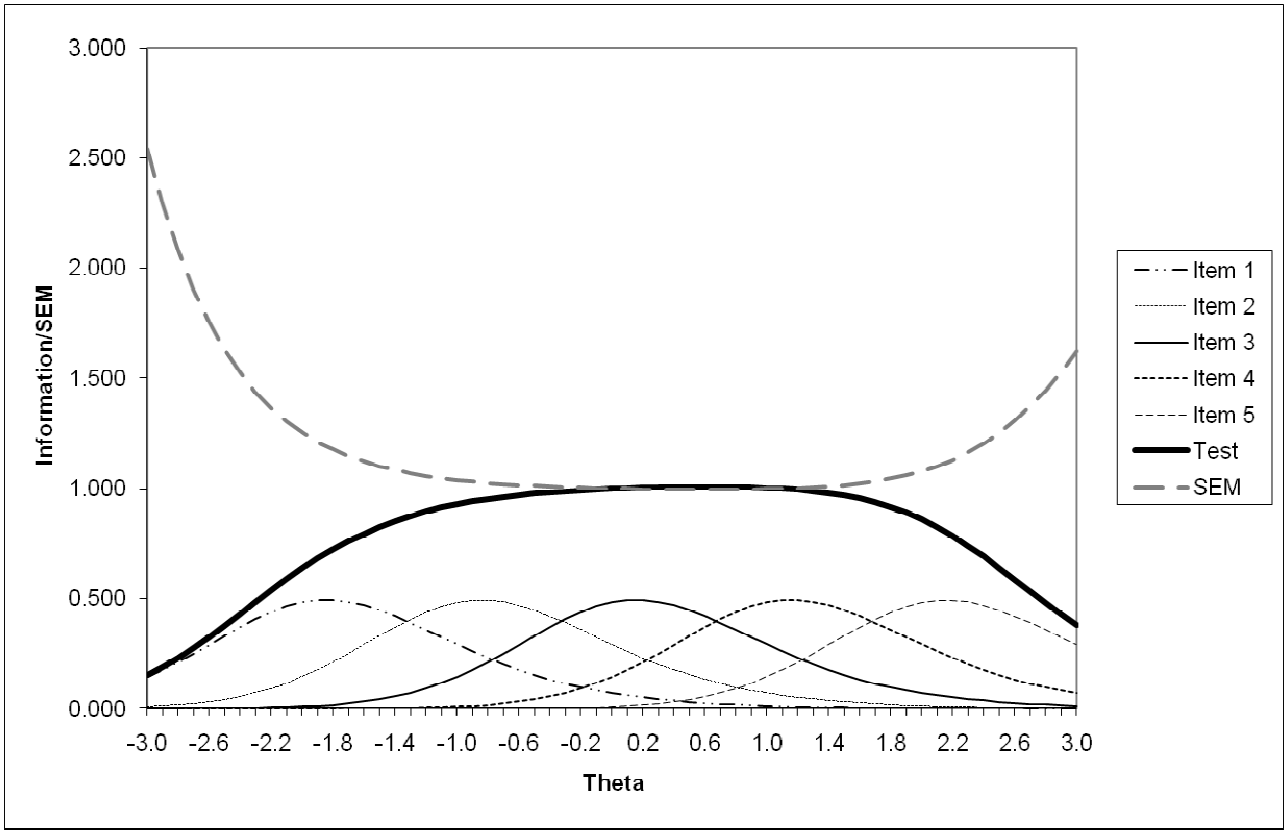


Figure 2

Item Information Functions, Test Information Function and SEM for a 5 item Test



References

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

Beaty, J. C., Grauer, E., & Davis, J. (2006). *Unproctored Internet Testing: Important questions and empirical answers.* Paper presented at the Practitioner forum conducted at the 21st Annual Meeting of the Society of Industrial and Organizational Psychology, Dallas, TX.

Bejar, I.I. (1983). Achievement testing: Recent advances. In J.L. Sullivan & R.G. Niemi (Eds.), Quantitative applications in the social sciences. (No. 07-036). Beverly Hills, CA: Sage.

Biskin, B. H., & Kolotkin, R. L. (1977). Effects of Computerized Administration on Scores on the Minnesota Multiphasic Personality Inventory. Applied Psychological Measurement, 1(4), 543-549.

Boyle, S. (1984). The effect of variations in answer-sheet format on aptitude test performance. Journal of Occupational Psychology, 57, 323-326.

Breithaupt, K. J., Mills, C. N., & Melican, G. J. (2006). Facing the opportunities of the future. In D. Bartram & R. K. Hambleton (Eds.), Computer-Based Testing and the Internet: Issues and Advances, pp. 231-251.

Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research and Practice, 33*, 148–154.

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the Internet. *European Journal of Psychological Assessment, 21*(2), 115–127.

Bugbee, A. C., Jr. (1996). The equivalence of paper-and-pencil and computer-based testing. Journal of Research on Computing in Education, 28(3), 282-99.

Byrne, B. M. (1998). *Structural equations modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*(3), 253-260.

Canoune, H. L., & Leyhe, E. W. (1985). Human versus computer interviewing. Journal of

Personality Assessment, 49, 103-106.

Drasgow, F., Nye, C.D., & Tay, L. (2010). Indicators of Quality Assessment. In J.C. Scott & D.H. Reynolds (Eds.), Handbook of workplace assessment: Selecting and developing organizational talent. San Francisco, CA: Jossey Bass, p XXX-XXX.

Equal Employment Opportunity Commission, Civil Service Commission, United States Department of Labor, & Department of Justice (1978). Uniform guidelines on employee selection procedures. Federal Register, 43(166), 38290-38315.

Fox, S., & Livingston, G. (2007). Latinos online: Hispanics with lower levels of education and English proficiency remain largely disconnected from the internet [Electronic Version], from http://pewhispanic.org/files/reports/73.pdf

Greaud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. Applied Psychological Measurement, 10, 23-34.

Guion, R.M. (1998). Assessment, measurement, and prediction for personnel decisions. Mahwah, NJ: Erlbaum.

Hambleton, R. K., & Pitoniak, M. J. (2002). Testing and measurement: Advances in item response theory and selected testing practices. In J. Wixted (Ed.), Stevens' handbook of experimental psychology (3rd ed., Vol. 4, pp. 517–561). New York: Wiley.

Hausknecht, J.P., Halpert, J.A., Di Paolo, N.T., Moriarty, N.T., & Gerrard, M.O. (2007). Retesting in selection: A meta-analysis of practice effects for tests of cognitive ability. Journal of Applied Psychology.

Horrigan, J. (2009). Home broadband adoption increases sharply in 2009 with big jumps among seniors, low-income households, and rural residents even though prices have risen since last year. [Electronic Version], from http://www.pewinternet.org/Press-Releases/2009/Home-broadband-adoption-increases-sharply-in-2009.aspx

International Test Commission (2006). International guidelines on computer-based and internet-delivered testing. International Journal of Testing, 6 (2), 143-171

Holland,W. H.,& Wainer, H. (1993). Differential item functioning. Hillsdale,NJ: Lawrence Erlbaum.

Jodoin, M.G. (Spring, 2003). Measurement efficiency of innovative item formats in computer-based testing. Journal of Educational Measurement, 40 (1), 1-15.

Kolen, M.J. & Brennan, R.L. (2004). Test Equating, Scaling, and Linking: Methods and Practices. New York, NY: Springer-Verlag.

Lievens, F., Buyse, T., & Sackett, P. R. (2005) Retest effects in operational selection

settings: Development and test of a framework. Personnel Psychology, 58, 981–1007.

Liu, C., Borg, I., & Spector, P. E. (2004). Measurement equivalence of the German job satisfaction survey used in a multinational organization: implications of Schwartz's culture model. *Journal of Applied Psychology, 89*(6), 1070-1082.

Mazzeo, J., & Harvey, A.L. (1988). The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature (ETS RR-88-21). Princeton, NJ: Educational Testing Service.

McPhail, S. M. (Ed.) (2007). *Alternative validation strategies: Developing new and leveraging existing validation evidence* . San Francisco: Jossey - Bass.

McPhail, S.M., & Stelly, D.J. (2010). Validation strategies. In J.C. Scott & D.H. Reynolds (Eds.), Handbook of workplace assessment: Selecting and developing organizational talent. San Francisco, CA: Jossey Bass, p 671-710.

Mead, A. D. (April, 2010). Non-comparability of speeded computerized tests: Differential speededness? Paper presented at the Annual Meeting of the Society of Industrial and Organizational Psychology, Atlanta, GA.

Mead, A. D., & Blitz, D. L. (April, 2003). *Comparability of paper and computerized non-cognitive measures: A review and integration* Paper presented at the annual meeting of the Society of Industrial and Organizational Psychology, Orlando, FL.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological Bulletin, 114(3), 449-458.

Mead, A. D., Segall, D. O., Williams, B. A., & Levine, M. V. (April, 1999). Multidimensional assessment for multidimensional minds: Leveraging the computer to assess personality comprehensively, accurately, and briefly. A paper presented at the twelfth annual conference for the Society for Industrial and Organizational Psychology, St. Louis, Missouri.

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007) Are internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. Organizational Research Methods, 10(2), 322-345.

Neuman, G. & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? Applied Psychological measurement, 22(1), 71-83.

Ones, D. S., & Viswesvaran, C. (1998). The effects of socially desirability and faking on personality and integrity assessment for personnel selection. Human Performance, 11, 245-269.

Outtz, J.L. (2010). Addressing the flaws in our assessment decisions. In J.C. Scott & D.H. Reynolds (Eds.), Handbook of workplace assessment: Selecting and developing organizational talent. San Francisco, CA: Jossey Bass, p 711-728.

Ployhart, R. P., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods*, 7, 27-65.

Paek, P. (August, 2005). Recent Trends in Comparability Studies. PEM Research Report 05-05. Pearson Educational Measurement.

Parshall, C. G., Davey, T., & Pashley, P. J. (2002). Innovative item types for computerized testing. In W. J. van der Linder & C. A. W. Glas (Eds.)

Pearlman, K. (2009). Unproctored Internet testing: Practical, legal, and ethical concerns. Industrial and Organizational Psychology, 2(1), 14-19.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. F. (2002, April). Web-based vs. paper and pencil testing: A comparison of factor structures across applicants and incumbents. Paper presented at the 17th annual conference of the Society for Industrial and Organizational Psychology, Toronto, CA.

Pomplun, M., Frey, S., Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. Educational and Psychological Measurement, 62, 337-354.

Rainie, L., Estabrook, L., & Witt, E. (2007). Information searches that solve problems [Electronic Version], from http://www.pewinternet.org/Reports/2007/Information-Searches-That-Solve-Problems.aspx

Raju, N. S., & Ellis, B. B. (2002). Differential item and test functioning. In Drasgow, F., & Schmitt, N. (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp.156-188). San Francisco: Jossey-Bass.

Raju, N.S., Laffitte, L.J., & Byrne, B.M. (2002). Measurement equivalence: A comparison of confirmatory factor analysis and item response theory. Journal of Applied Psychology, 87(3), 517-529.

Raymond, M. R., Neustel, S., & Anderson, D. (2008, March). The benefits of taking an identical version of a certification test on two occasions. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New York, NY

Reynolds, D.H., & Rupp, D.E. (2010). Advances in Technology-Facilitated Assessment. In J.C. Scott & D.H. Reynolds (Eds.), Handbook of workplace assessment: Selecting and developing organizational talent. San Francisco, CA: Jossey Bass, p XXX-XXX.

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. Journal of Applied Psychology, 84(5), 754-775.

Rogers, J., Howard, K., & Vessey, J. (1993). Using significance tests to evaluate equivalence between two experimental groups. Psychological Bulletin, 113(3), 553-565.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. Journal of Applied Psychology, 83(4), 634-644.

Ryan, A. M., Horvath, M., Ployhart, R. E., Schmitt, N., & Slade, L. A. (2000). Hypothesizing differential item functioning in global employee opinion surveys. Personnel Psychology, 53(3), 531-562.

Sacher, J., & Fletcher, J. D. (1978). Administering paper-and-pencil tests by computer, or the medium is not always the message. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference (pp. 403-419). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Segall, D. O. (1996). Multidimensional adaptive testing. Psychometrika, 61, 331-354.

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

Stanton, J. M., & Rogelberg, S. G. (2001). Using internet/intranet web pages to collect organizational research data. *Organizational Research Methods, 4*, 199-216.

Stark, S., Chernyshenko, O. S. & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., et al. (2006). Unproctored Internet testing in employment settings. Personnel Psychology, 59(1), 189-225.

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.

Weiss, D. J. & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. In D. J. Weiss (Ed.). Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing. Minneapolis: University of Minnesota, Department

of Psychology, Psychometric Methods Program.

Weiss, D.J. (2007). Adaptive—and Electronic—Testing: Past, Present, and Future. Invited address presented at the annual meeting of the National Council on Measurement in Education, Chicago IL.

Zenisky, A. L.,&Sireci, S. G. (2002). Technological innovations in large-scale assessment. Applied Measurement in Education, 15(4), 337–362.

Zickar, M., Rosse, J., & Levin, R. (1996, April). Modeling the effects of faking on personality instruments. Paper presented at the 11th annual meeting of the Society for Industrial and Organizational Psychology, San Diego.