Assessment Selection Standards Guide

Dennis W. Koerner, Ph.D.
Russell J. Watson, Ed.D.

Table of Contents

Background

Assessments are ubiquitous in today's business world, with widespread availability and use. With the advent of the internet in the 1990s, the ability to produce, market, and sell assessments became exponentially easier and less expensive. Since then, it has developed into a kind of "global cottage industry" with hundreds of new assessment developers producing thousands of different assessments. Assessments are now available for almost any type of psychometric measure imaginable. A few simple examples include, personality, behavior, emotional intelligence, leadership, management potential, motivation, driving forces, critical thinking, reliability, reasoning, and so on. To further illustrate the point, the Society of Human Resources Management (SHRM) has its own Talent Assessment Center (TAC) for companies to purchase assessments. The TAC catalog indicates that there are 427 tests and assessments available for sale. More specifically, there are 182 assessments under the category of "Personality" assessments alone. This is just one source of assessments and there are many more.

Psychometric tests are widely used because they are a cheap and effective way of distinguishing between candidates and accurately identifying who is likely to be successful in the job role. They can be administered to candidates early on in the process and don't require a face to face meeting, thereby reducing the time and costs associated with selection. This cost savings and ease of use has resulted in the widespread proliferation of assessments. As a result, users are now faced with two exceedingly difficult problems:

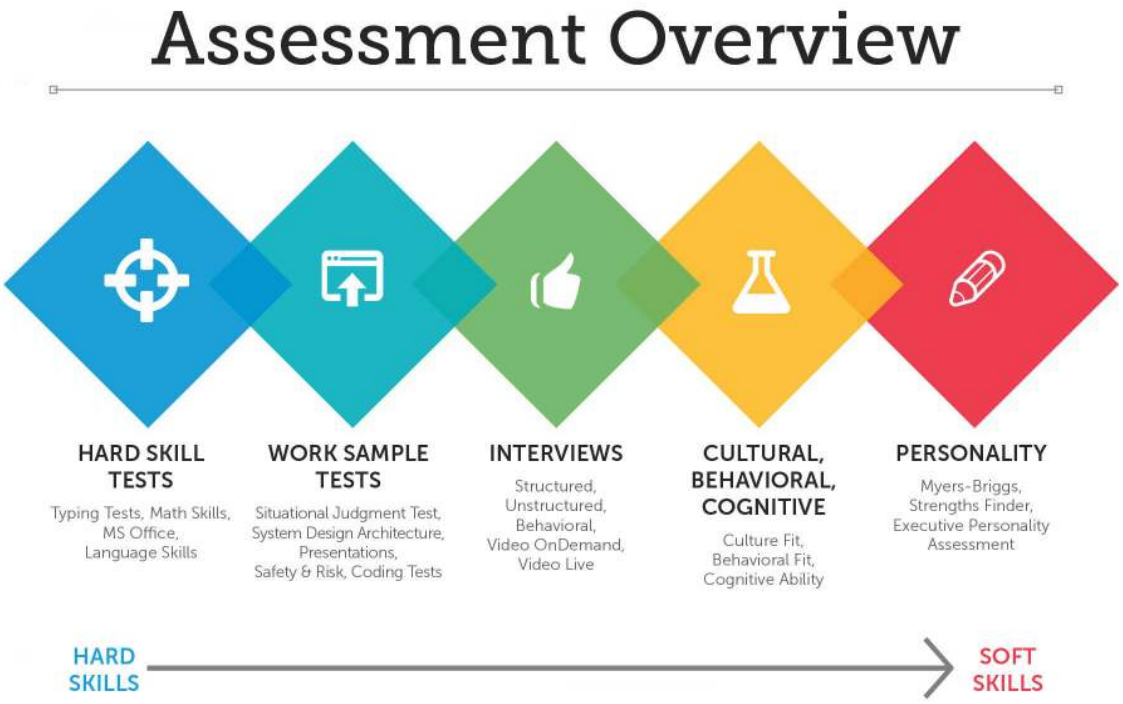
(1) ***With so many assessments available, how do you know which assessment to use?***

(2) ***How do assessment users know if the assessments they are using work as advertised?***

Assessment Standards

Assessments were initially produced by only a handful of academically trained and skilled Ph.D. psychologists. Today, assessments can be produced by anyone. No credentials or understanding of assessment design principles are required. Fortunately, numerous professional organizations have established standards for testing and assessments. The most significant of these are the American Psychological Association (APA) and the Equal Employment Opportunity Commission (EEOC). We strongly encourage that APA standards for Data Stability, Data Reliability and the various types of validity be adopted and used by the industry. Even more important, practices for ensuring no Disparate Impact should follow the EEOC guidelines.

The purpose of this article is to provide you with a review of the tools essential to selecting the appropriate reliable, valid, and unbiased assessments that are suitable for your assessment needs.



Theoretical Basis

The first step in the process of assessment selection is to ensure that the test construct is suitable for your use. For this you need a theory. Assessment theory provides an explanation of the purpose of the assessment and its subsequent question rationale. A well thought and proven theoretical basis is essential to explaining your assessment findings and making subsequent judgements.

Here are a few questions designed to help you determine if there is a sound theoretical basis for the assessment you choose.

- Is the purpose of the Assessment specified?
- Is the assessment theory described?
- Does the assessment state what it will measure?
- Are applicable target group factors, such as age, gender, education level, and other relevant factors specified?
- Is the assessment job-relevant?

Data Reliablity

Assessment standards require dependable measurement. Measurements are deemed to be reliable to the extent that they are repeatable and that any random influence which make measurements different from one occasion to another is a source of measurement error. Reliability is the degree to which an assessment consistently measures whatever it is supposed to measure. There are two general types of widely used reliability measures:

- **Test Stability:** Test-retest reliability is the degree to which scores are consistent over time. In other words, if the assessment were given again, would the results produced be the same?

- **Internal Consistency:** Internal consistency evaluates individual questions in comparison with each other for their ability to provide consistently appropriate results. Internal consistency measures whether several items (such as a list of questions) that propose to measure the same general attribute produce similar scores.

To help you select an assessment with good data reliability you should ask for the following reliability data:

- Is there information about the reliability of the assessment?
- Is the data stable over time?
- What is the test-retest reliability?
- Does the internal consistency measure meet American Psychological Association (APA) standards?

Validity

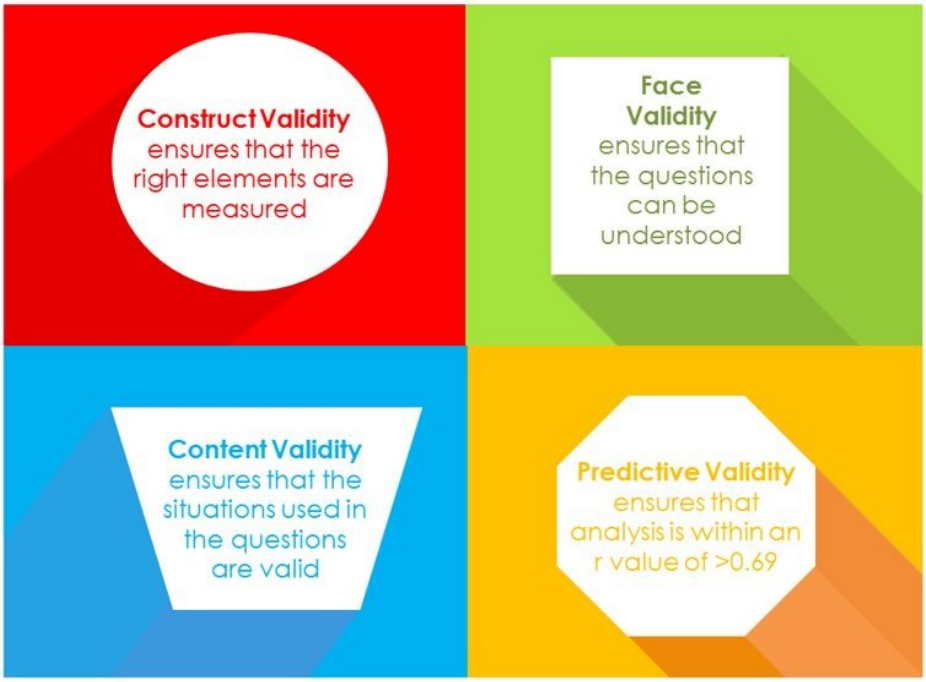
The validity of an assessment is the degree to which it measures what it is supposed to measure. There are many types of validity measures. The four main categories of validity used by the Assessment Standards Institute in evaluating assessment instruments are: Face Validity, Content Validity, Construct Validity and Criterion Validity.

Many assessments claim validity but few assessment analysis reports define the type of validity or state its actual measure. A brief description of each validity type is given below.

- **Face Validity** is an important consideration in both instrument construction and reporting when being used in a business setting. Face validity considers whether the questions appear relevant to the respondent and whether the report findings appear accurate to the reader. That is, does the assessment appear to measure what it aims to measure? This is the least sophisticated measure of validity.
- **Content Validity** uses recognized subject matter experts to evaluate whether test items assess if the assessment reflects the knowledge required for a given topic area.
- **Construct Validity** examines the ability of an instrument to measure a theoretical construct or trait. This form of validity is built from multiple measures and a pattern of evidence across several sources. Further construct validity can be determined by examining correlations among assessments.
- **Criterion Validity** is the extent to which an assessment measure is related to an outcome. In this case the assessment is predicting a measured dependent variable outcome.

Questions to help you discern assessment validity:

- Is there information related to the validity of the assessment?
- What type of validity do you need?
- What is the quality of the validity data?



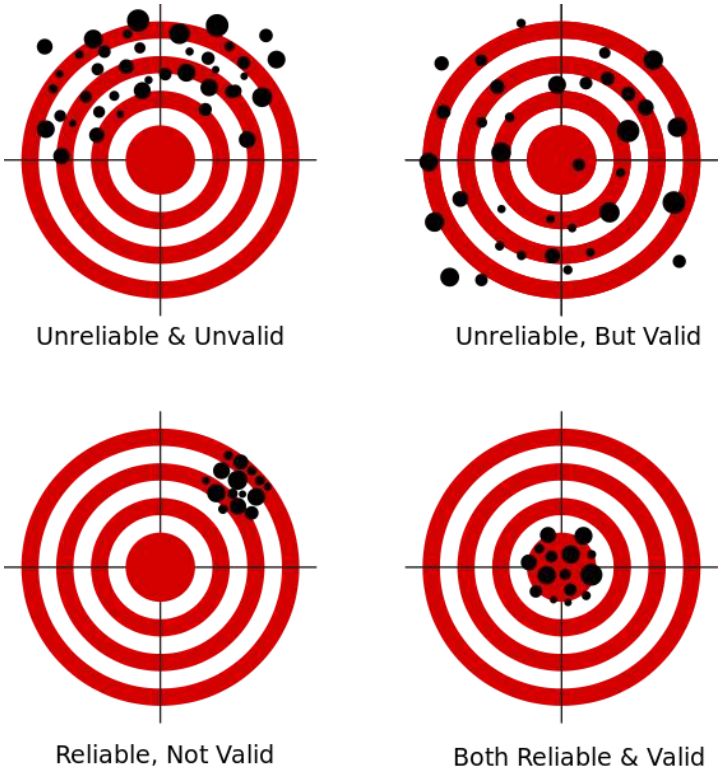
It is important to remember that primary to a test construct and usefullness is data reliability. Many assessments claim to have validity, however, often the type of validity claimed is not mentioned. When evaluating an assessment, be sure that the assessment data is reliable and know what type of validity is being claimed.

Data Reliability vs Data Validity

In the last two sections we have learned the importance of having both the correct type of data reliability and validity. It is essential to understand the difference between the two measures. To help us get a better perspective of the subtleties, let's follow a simple example of going to the doctor to have your temperature measured.

Data Reliability - The physician measures your forehead temperature several times under identical conditions. The thermometer displays the same temperature every time, so the results are reliable. If, however, the thermometer displays different temperatures, then the results are deemed to be unreliable.

Data Validity - If the physician's thermometer is calibrated and reads the true temperature of your forehead, the result is deemed to be valid. If, however, the thermometer is not calibrated but provides a consistent reading, the reading is not the actual temperature. and this result is deemed to be invalid.



Disparate Impact

The Equal Employment Opportunity Commission (EEOC) has established specific guidelines for the use and implementation of pre-employment assessments. These guidelines include specific statements regarding the avoidance of discriminatory bias (Disparate Impact) against any EEOC protected groups. There are generally two guidelines that are applied. The procedures below describe how an employer can defend a process that has been identified as having adverse impact. Adverse impact can occur when identical standards or procedures are applied to everyone, even though they lead to a substantial difference in employment outcomes for the members of a particular, protected class.

According to the EEOC Guidelines:

"Each user should maintain and have available for inspection records or other information which will disclose the impact of its tests and other selection procedures"

What You Need to Know

As an important note, the presence of disparate impact does not require the elimination of the procedure (e.g. assessment use), but rather its required justification as being job-related or a business necessity.

EEOC Required Information

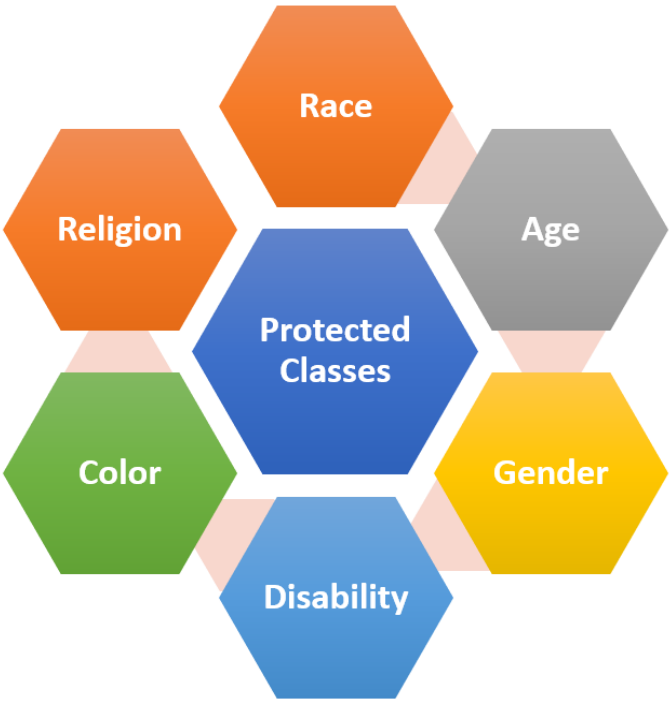
- **The 80% Rule (for small data sets)** - Typically, for small data sets, Disparate Impact is determined by using the four-fifths or eighty percent rule. The four-fifths or 80% rule is described by the guidelines as "a selection rate for any race, sex, or ethnic group which is less than four-fifths (or 80%) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of disparate impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of disparate impact." Since the 80% test does not involve probability distributions to determine whether the disparity is a "beyond chance" occurrence, it is usually not regarded as a definitive test for adverse impact. Instead, other statistically significance tests, may be used for this purpose.

- **Statistical Significance (for large data sets)** - A Z-test of independent proportions, often called the 2 standard deviation test, is a statistical technique that translates the probability of a difference in selection rates into the metric of standard deviations. Many court cases as well as the OFCCP have adopted the enforcement standard of two or more standard deviations as an indication of statistical significance in employment discrimination cases.

Questions to help you discern conformance to EEOC guidelines:

- Does the assessment have a Disparate Impact data analysis and report?
- What standards are used to establish compliance?
- Were the appropriate protected class comparisons made?
- What is the measure of statistical significance?

EEOC Protected Classes



Data Standards

Test results are usually presented as a “raw score” per assessment user. To establish compliance to standards, data from the raw scores are subsequently aggregated and analyzed. The analyzed data when compared to standards, then determines the various properties and usefulness of the assessment. Key standards for each of the measures listed above should be referenced by the author of the assessment.

Resources for Standards

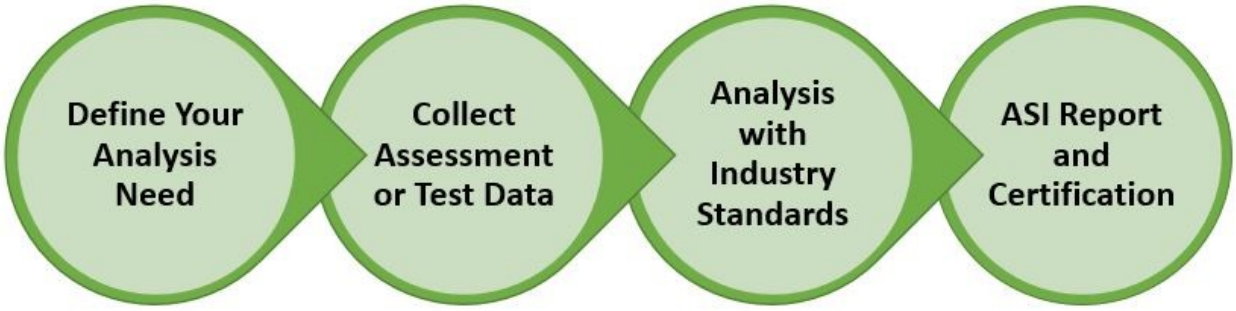
Questions to help you determine the quality of your assessment:

- Are data standards published for each test result and analysis?
- Does the assessment meet the minimum requirements of the APA and EEOC?
- How do other assessment results compare?

Third Party Analysis

It is often found that companies make claims related to their assessment data but do not actually have the data necessary to substantiate the claim. Given the detailed, mathematical nature of data analysis, some analysis is conducted “in house” and found to be inaccurate. As with most “verified” analysis, customers typically require an outside, skilled, third-party to conduct the analysis and verify the results.

Third-party analysis and certification means that an independent organization has reviewed various aspects of the data associated with the assessment and has independently determined that the final product complies with specific standards for use, quality or performance.



The process for evaluating your test or assessment is straightforward and simple. The process begins by determining what analysis is need. The classic approach is to have data analyzed for (1) Internal Data Reliability (2) Construct Validity and (3) Disparate Impact. Once you have determined your needs, collect your data and submit to a qualified, third party vendor. The third party should then provide you with a report and applicable dates.

What is an applicable date for certifications? Words and understanding of words change with time. It is important to update the analysis of your tests or assessments to assure their standards compliance.

Take a Holistic Approach

When using your assessment, please realize that assessments are not clinical tools that have great predictive powers. Rather they provide small and valuable insights that help you to better communicate and understand another person. Using a single test or procedure will provide you with a limited view of a person. Moreover, you may reach a mistaken conclusion by giving too much weight to a single test result. On the other hand, it is well established that using a variety of assessment tools enables you to get a more complete picture of the individual. A holistic view of a person is the best view.

Above is a diagram of some factors you may want to consider as part of your assessment analysis. Realize that assessments typically only relate to one aspect of a person such as their behaviors, values or driving forces. Consideration of these attributes in light of other factors such as skills, prior experience, general mental ability and emotional intelligence will provide you a greater context in which to evaluate the person and your assessment results.

The Value of Assessments

In today's business world, psychometric assessments are widely used at every stage of an organization's talent management process, starting from talent acquisition to talent development. Psychometric assessments greatly enhance the chances of organizational success by ensuring that the right fit candidates are hired, identified and developed for critical roles. Whether it is employee engagement, appraisals, training needs identification, improving communication, leadership development, or succession planning, uses of psychometric tests can be found throughout the employee life cycle.

Like other management tools, assessments can be extremely helpful when used properly, but counterproductive when used inappropriately. Often inappropriate use stems from not having a clear understanding of what you want to measure and how you want to measure it. As outlined above, the first principles of assessment use are:

(1) Be sure to have a clear understanding of what needs to be measured and for what purpose.
(2) Use appropriate data standards to select the assessment that will best meet your needs.

Before you use or develop assessments, it is important to make sure that your measurement tools will provide you with unbiased, stable, reliable and valid data.

ASI | Assessment Standards Institute

The Assessment Standards Institute has the tools you need to ensure the success of your assessments.

The Assessment Standards Institute

The Assessment Standard Institute is an organization dedicated to the helping others to develop and use assessments in a legally compliant and effective manner. The process by which we achieve our mission is through the services we provide our customers.

Certification: Certify your tests and assessments. Know that they meet government regulations and the highest of industry standards.

Advisory: We'll help you to identify new opportunities, accelerate product development, access new markets, comply with regulatory requirements, and achieve your key business objectives.

Contact Us
For More Information

website: assessmentinstitute.org
email: dennis.koerner@assessmentinstitute.org
phone: 901.568.3569