Reliability and Validity in Psychometric Assessments for Hiring Decisions

Robert Stokes

November 15, 2024

1 Introduction

Within the assessment industry, there is a notable discussion surrounding the concept of validity coefficients. While reliability coefficients are well-established and widely accepted in assessment practices, the idea of a singular "validity coefficient" is not similarly recognized. This distinction stems from the understanding that validity cannot be encapsulated in a single numerical figure; doing so would significantly oversimplify the complexities involved.

Validity is better understood as a comprehensive collection of evidence that collectively demonstrates an assessment's ability to accurately measure the intended structured concepts, or *constructs*, fulfill its claims, and function as intended in practical applications. This nuanced perspective on validity highlights the importance of a multifaceted approach to evaluating assessments, ensuring their effectiveness and integrity across various contexts.

In the context of hiring, psychometric validity takes on an even more complex role. When deciding how much importance to place on psychometric assessments in hiring, it is recommended that they make up no more than 33% of the decision. These tests should be balanced with the candidate's qualifications and a formal interview. Psychometric assessments are useful for identifying personality traits, emotional intelligence, and soft skills, which help in evaluating how well someone fits into the company culture and the role for which they are applying.

2 Reliability and Validity Overview

Reliability is defined by the American Psychological Association (APA) as "the trustworthiness or consistency of a measure, that is, the degree to which a test or other measurement instrument is free of random error, yielding the same results across multiple applications to the same sample." Essentially, this means that an assessment should give consistent results each time it is used with different respondents under similar conditions. There are various methods to measure reliability, and each helps build a complete understanding of how consistently

the assessment measures the mechanism intended for measure. The APA has identified four major areas of focus for reliability study in *APA Handbook of Testing and Assessment in Psychology Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, which are described below.

- Temporal Consistency: "The correlation between measurements obtained from the same test or instrument when administered to the same sample on two different occasions. Temporal consistency is an index of the retest reliability of an instrument. It assumes that there is no substantial change in the construct being measured between the two occasions. The longer the time gap, the greater the likelihood of a lower correlation." Also known as test-retest reliability, this is a measure of how the assessment measures the stability of assessment responses over time. This type of reliability assesses whether individuals receive similar scores when they take the same assessment on two different occasions, assuming that no substantial changes have occurred in what is being measured. A high level of test-retest reliability suggests that the assessment can accurately reflect stable traits or abilities, as opposed to temporary states or momentary fluctuations. Typically, this is measured with by correlating the results of the same individuals between assessments taken in separate time periods.
- Alternate-Forms Reliability: "A measure of the consistency and freedom from error of a test, as indicated by a correlation coefficient obtained from responses to two or more alternate forms of the test." In this examination of reliability, two forms of the same assessment are generated and provided to the same population. This method attempts to show that the items on assessment A function in the same way as the items on test form B, in much the same way as in temporal consistency.
- Internal Consistency: The APA defines internal consistency as "the degree of interrelationship or homogeneity among the items on a test, such that they are consistent with one another and measuring the same thing." This form of reliability assesses that the items themselves function consistently across different individuals and consistently measure the same thing.
- Generalizability Theory: "A framework of principles and assumptions about how to determine the reliability of a set of data. Researchers investigate the various facets of a study (items, raters, settings, etc.) to understand specific sources of error and to determine the conditions under which observations will be consistent and applicable across different contexts (e.g., age groups, geographic regions, socioeconomic status)." By far, this is the most complex and difficult of the mentioned reliability measures, but no less important to consider. In developing and examining these assessments, aside from determining the level of error or variation in the assessment results, it is important to discover, or account, for the sources of the variation itself. This method focuses on separating poten-

tial sources of variation, or *facets*, and assessing their effect upon the data itself.

Validity as a concept has much greater breadth and depth. At its core, validity is a composite of the measure of many different aspects answering a single question: "Does this assessment do the things that it claims to do?" or, as described by definition from the APA: "The extent to which evidence and theory support specific interpretations of test scores for the proposed use of the test." Validity has multiple forms, depending on the research question and on the particular type of inference being made. A close examination of the above definition pulls together the following idea: There is no single thing that is a "Validity Coefficient". Validity is established by the repeated, plentiful and rigorous application of scientific methods to uphold the statements made by the assessment. More simply put, validity is measured in volume of supporting evidence, not a single coefficient.

3 Understanding Validity in Detail

Validity is confirmed by aligning the dimensions of measurement (the constructs) with the intended purpose of the assessment.
Below are five dimensions of validity discussed in psychometric assessments (as defined by the APA, then further explained):

- Test Content Validity: "The extent to which a test measures a representative sample of the subject matter or behavior under investigation." In order for an assessment to be valid, it must accurately represent the specific traits or abilities it claims to measure. Additionally, it's essential that the assessment avoids measuring any qualities or skills outside the intended constructs. This focus ensures that the assessment's results are relevant and meaningful, reflecting only the characteristics it is designed to evaluate. This method achieves this goal by examining the theoretical constructs and ensuring that the items within the assessment align properly.
- Internal Structure Validity: This form of validity refers to how well the items work together to measure the constructs within an assessment. Put into more detail, it is a measurement of how well the items represent the underlying factor structure of the assessment. As a clarifying example, an investigator would examine all of the Dominance items on the TTISI DISC assessment and determine how well those items interact with one another and separate themselves from the I,S, and C items.
- Evidence of Validity in Relation to Other Variables: This form of validity refers to a study into how the assessment constructs relate to variables and other constructs outside of the assessment. This is often the most difficult form of validity to use in building a strong validation argument. The difficulty is that the external criterion must be heavily scrutinized for

practicality, relevance, and reliability. For example, any construct within the TTISI DISC assessment would likely find little utility by including a comparison between items and the measurement of oceanic salt content. However, there may be justifications to examine the connection between DISC variables and particular O*Net job classifications, as some jobs collect, to a degree, predictable DISC profiles.

- Consequences of Testing Validity: According to *APA Handbook of Testing and Assessment in Psychology Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, "Validity evidence based on consequences of testing refers to evaluating both the intended and the unintended consequences associated with a testing program." This type of validity is often the most difficult type to produce. It relies on the individuals physically administering the assessment to utilize caution and intention with respect to the intended purpose and consequences of utilizing the assessment itself. Putting it simply, in order to adequately measure this form of validity, there must be communication about the utilization of the assessment and evidence that it is being used properly and without unintended results. As an example, the assessment should not be used in such a way that leads to a discriminatory practice.
- [Validity] Evidence Based on Response Processes: "Some construct interpretations involve more or less explicit assumptions about the cognitive processes engaged in by test takers. Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers." In response-process validity, the goal is to analyze the specific steps a respondent follows to answer test questions accurately, assessing if the process itself aligns with the intended measure. This approach presents unique challenges in psychological or cognitive assessments, as the process of responding often directly reflects the very cognitive abilities or psychological traits under evaluation. TTISI sets itself apart in the assessment industry in this way by incorporating evidence from direct EEG measurements of brain activity captured while respondents interact with and complete test content. This direct observation of brain responses to specific stimuli provides valuable insights into the cognitive processes at play and showing the innate connection between process and response.

The above list is by no means a complete list of validity forms, but serves to demonstrate a strong counter-argument against the misconception that validity can be represented by a single coefficient. Additionally, these examples demonstrate that validity involves a rigorous accumulation of evidence across multiple categories.

4 Practical Application in Hiring

The question of how much weight should be given to psychometric assessments in the hiring process requires careful consideration. These tools, by their own nature, are limited to assessing the specific constructs they are designed to measure. While valuable, psychometric assessments can be influenced by external factors such as a candidate's mood or interpretation of the assessment questions, which may affect the results. The key to their utility lies in selecting the appropriate assessments or suite of assessments for the specific hiring needs at hand. Given these limitations, psychometric assessments should account for no more than 33% of the overall hiring decision process.

5 Conclusion

Validity is a complex, multi-dimensional concept that is not reducible to a single statistic. It is built on a collection of evidence, showcasing how an assessment measures its intended constructs, relates to theoretical frameworks, and serves its intended purpose. When used correctly, psychometric assessments can play a valuable role in hiring, but they must be used in conjunction with other tools to ensure well-rounded decision-making. For further information on the validity of specific assessments, readers should refer to comprehensive guides such as the Workplace Competencies Technical Manual (https://ttiresearch.com/project/workplace-competencies-technical-manual-version-1-0/).

For further reading on validity, please see:
American Psychological Association Guidelines on Psychometric Assessment Validity and Reliability Requirements, Gehrig, E, 2019.
TTI Success Insights Approach to Psychometric Assessment Validity and Reliability, Gehrig, E, Bonnstetter, R, 2019
Some Thoughts on Current Consensus Views on Evidence of Reliability and Validity in the Psychometric Assessment World, Gehrig, E, Bonnstetter, R, 2019
Standards for Educational and Psychological Testing, Chapter 1, American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for educational and Psychological Testing (US), 2014
Brain Activation Imaging in Emotional Decision Making and Mental Health: A Review—Part 1, Bonnstetter, R, Collura, T, 2021, Clinical EEG and Neuro-science, 52(2), 98-104.

For more information about assessments contact The DISC Agency on 1300 690 469 or email support@thediscagency.com.au www.thediscagency.com.au