CEBMa
center for
Evidence-Based Management

Psychometric Concepts:
A Practical Guide

November 2025

Introduction

An important part of evidence-based decision-making is measurement. In the first step, you measure variables to understand the problem and its seriousness. At the end, you assess outcomes to determine whether the intervention, practice, or policy actually solved the problem. Often, several tools—scales, tests, questionnaires, and other instruments—are available. The question, however, is how trustworthy these tools are.

To judge whether an assessment tool is trustworthy—that is, valid and reliable—it helps to understand the basics of psychometrics, the discipline concerned with measuring mental capabilities, traits, attitudes, and perceptions. To empower evidence-based practitioners to make informed decisions, it helps to know what constitutes a reliable and valid tool, how to avoid common pitfalls, and the right questions to ask consultants advocating for a particular tool. This guide summarizes the most relevant psychometric concepts to consider when evaluating the validity and reliability of assessment tools.

Questionnaires, Scales, and Items

Assessment tools take various forms, such as questionnaires, online tests, or structured interviews. Questionnaires typically consist of a series of questions or statements presented to respondents. In academia, questionnaires are sometimes referred to as 'scales,' with questions or statements referred to as 'items.' Although these terms are used interchangeably, a scale is a specific type of questionnaire that focuses on measuring a particular construct or trait, such as 'organizational commitment' or 'self-esteem'. Questionnaires are commonly used to collect more general types of information, such as a person's satisfaction with a service or product.

In both cases, a 'rating' scale is often used – sometimes referred to as a "Likert scale" after the psychologist who first created it - asking respondents to indicate their level of agreement or disagreement with each statement. This way responses can be averaged to create a numerical score for each respondent, allowing for quantitative analysis.

Standardization

A test or questionnaire is *standardized* when it is given and scored in the same way for everyone. This means:

- the items are fixed and appear in the same order
- everyone gets the same instructions
- everyone uses the same response options (for example, a 1–7 Likert scale)
- the scoring method is the same for all respondents

In general, there are two types of standardization:

1. Norm-referenced standardization

Some tests (e.g., IQ or aptitude tests) include norms or cut-off scores that allow you to compare a person's score to a reference group. These tests often have strict rules and defined ranges for what counts as "high" or "low."

2. Procedural standardization

Most questionnaires in organizations use this form. The focus is on consistency of administration and scoring. No time limits or norm tables are needed. This type of standardization is the most common.

Types of Assessments

When measuring variables and outcomes, we distinguish between **people metrics**—human attributes, behaviors, or perceptions—and **operational metrics** based on hard data such as revenue, costs, turnover, absenteeism, and safety incidents. For the latter, psychometrics are irrelevant and therefore fall outside the scope of this guide. People metrics are commonly grouped into the following types of assessments:

1. Individual-level assessments

Tests that measure *attributes* of a person: cognitive ability, aptitude, personality, skills, preferences. Used for selection, development, placement.

2. Organizational-level measurement scales

Surveys measuring *perceptions* of individuals about a construct: engagement, commitment, climate, culture, psychological safety, leadership style.

3. Performance ratings

Assessments in which supervisors, peers, or subordinates rate a person's behavior or performance, such as 360-degree feedback or competency ratings. They are not tests, but a type of assessment that still requires reliability and validity.

Relevant Psychometric Properties of Assessment Tools

The quality of an assessment tool can be determined by the psychometric properties of the questionnaire or scale it uses. Two key psychometric properties are reliability and validity. However, before determining a tool's reliability and validity, it should first be clear what its purpose is.

A. Purpose

Based on the information provided by the developer of the assessment tool, users should be able to judge whether it fits the purpose for which they are seeking a tool. Therefore, a clear description of the tool's purpose, target group, content, and conditions of use should be available. Relevant questions to consider include:

1) *Is the purpose of the tool clearly stated?*
2) *What construct(s) is the tool intended to assess?*
3) *What specific outcomes (behavior(s) or performance) is the tool intended to predict?*
4) *What is the target population the tool assesses?*
5) *Are the instructions on how to use the tool complete and clear?*
6) *Are the instructions on how to interpret the results detailed and clear?*
7) *Are the limitations of the tool's usability and outcome(s) clearly specified?*
8) *Is information provided on possible differences in outcomes between subgroups (e.g., gender, educational level, years of experience, ethnic background)?*

In some cases, the assessment tool may be administered by a third party other than the HR professional (e.g., a vocational psychologist), in which case questions 5 and 6 may be less relevant.

B. Reliability

The reliability of an assessment tool indicates whether it produces the same outcome when administered at different times under the same circumstances. For instance, if a scale[1] measuring personality traits indicates that an employee scores high on extraversion, the scale should yield the same result when administered at another time. There are several properties that determine whether a scale is reliable, the most relevant are:

1. Stability

An important property of a scale is whether the outcome remains stable over time. This can be determined by comparing the results at two different points within a relatively short period of time. This property is referred to as test-retest reliability. Test-retest reliability coefficients range from 0 to 1, where a score of .60 is considered acceptable and values above .80 good.

2. Internal Consistency

Internal consistency, also referred to as homogeneity or inter-item agreement, is a reliability property that refers to whether items of a scale measure the same characteristic. For example, if an assessment tool consists of a scale of which three items measure the concept of 'resilience', then individual respondents should consistently score similarly on all three items. Internal consistency is typically evaluated using measures such as:

[1] Most people assessment tools take the form of scales. For this reason we use the term 'scale' interchangeably with 'assessment tool', emphasizing the relevance of psychometric properties.

- ***Split-Half Reliability***

Split-half reliability involves splitting a scale into two halves and then comparing the scores to evaluate the consistency of measurement. Splitting a scale can be done in various ways, such as splitting odd- and even-numbered items, dividing the scale into two equal parts, or using random assignment. A common guideline is to consider a split-half reliability coefficient of .70 as acceptable, with values above .80 as good.

- ***Cronbach's Alpha and Composite Reliability***

Cronbach's alpha is the most widely used measure for internal consistency. There is, however, no consensus on its interpretation. While some studies propose that values above .70 are ideal, some researchers find values close to .60 satisfactory. The measure of composite reliability is much like Cronbach's alpha and is interpreted in a similar way: higher values indicate better internal consistency.

- ***Inter-item Correlation and Item-to-scale Correlation***

While Cronbach's alpha and composite reliability capture the overall internal consistency of an assessment tool, inter-item and item-to-scale correlation examine the relationships between the items. For example, they assess whether a score on one item is related to the scores on other items (inter-item correlation), or whether a score on an individual item is related to the overall score of the entire scale (item-to-scale correlation). Ideally, the inter-item correlation should be between .20 and .40, whereas the item-to-scale correlation should be above .30 (acceptable) or .50 (strong).

3. Equivalence

When an assessment tool requires human judgment, the degree of agreement among two or more raters is an important measure for the tool's reliability. This property is referred to as Interrater reliability. Values above .70 are often considered acceptable, while values above .80 indicate good reliability.

C. Validity

The validity of an assessment tool indicates whether the scale(s) it uses measures what it is supposed to measure. In the research literature, validity takes a variety of forms. However, for people assessment tools, the three most relevant properties are content validity, construct validity, and criterion validity.

1. Content Validity

Content validity refers to the degree to which the content of a scale (its items) adequately reflects the construct being measured. To produce a valid outcome, the items of a scale must cover all elements of the construct it aims to measure. If some aspects are missing or irrelevant aspects are included, the tool has low construct validity. For example, if a scale aims to measure employee performance but items regarding punctuality, adaptability and collaboration with team members are missing, it has low content validity. Since there is no

statistical test to assess content validity, researchers typically rely on the judgment of experts. Because content validity is a subjective measure, it does not provide conclusive evidence of the scale's validity. Note that related but distinct from content validity is what is called face validity, that is, a quick informal impression of whether the scale "looks like" it is appropriate and relevant to its conditions of use. Face validity typically refers to the judgments made by respondents to whom the scale is administered.

2. Construct validity

Construct validity refers to the extent to which a scale assesses the construct it is intended to measure. Put differently, does the scale measure the intended construct or, in part or primarily, something else? For instance, if a scale that aims to assess the construct of general intelligence measures shoe size, that scale has low (or, more accurate: zero) construct validity, as shoe size and intelligence are not related. The term 'construct' refers to a characteristic or concept that can't be directly measured but can be assessed through indicators associated with it. For example, the construct 'intelligence' can't be measured directly, but we can measure indicators of intelligence, such as analytical thinking, problem-solving skills, and learning ability. This is why a scale typically uses multiple items to measure a construct. Establishing construct validity is challenging. It typically requires multiple studies and involves accumulating research results. Additionally, a solid theoretical framework is crucial. As a result, construct validity cannot be fully captured by a single score or indicator; it reflects multiple properties and indicators, such as:

- ***Convergent validity***

Convergent validity reflects the extend to which the scores of two scales that measure the same construct are related. For example, if a scale measures job satisfaction, its scores should be highly correlated with the scores of other scales measuring job satisfaction.

- ***Discriminant validity***

Discriminant (or divergent) validity reflects the extent to which the scores of two scales that measure different constructs are related. For instance, if a scale measures job satisfaction, its scores should show only weak or no correlation with the scores of a scale measuring employee extraversion or proactivity. Discriminant validity is essential to ensure that a scale does not inadvertently measure a different construct. For instance, the Q^{12} Survey developed by Gallup claims to assess employee engagement. However, the Q^{12} scale exhibits a nearly perfect correlation ($r = .91$) with a single item measuring job satisfaction (Harter et al, 2002). This suggests that the Q^{12} may measure job satisfaction rather than employee engagement, indicating a low level of construct validity.

- ***Structural/factorial validity***

Structural or factorial validity refers to the degree to which scores on a scale accurately represent the underlying dimensions (factors) of the construct it is designed to measure. Factor analysis, a widely employed statistical method, is commonly used to assess this property. Researchers use factor analysis to determine whether the items in a scale

effectively capture the underlying structure and dimensions (factors) of a construct, as well as whether the items sufficiently differentiate between related but distinct constructs. For example, the NEO Personality Inventory (NEO-PI), a widely used tool for assessing an individual's personality, measures five personality traits with a total of 240 items. Due to the extensive number of items, a shorter version, the NEO Five Factor Inventory (NEO-FFI), consisting of only 60 items, was developed. However, several studies using factor analysis found that some items in this shortened version capture more than one trait (Egan, 2000), suggesting that the shorter NEO-FFI has lower construct validity than the longer NEO-PI

3. Criterion validity

Criterion validity reflects the extent to which a scale is related to an external criterion. This external criterion should be a widely accepted measure, typically an objective outcome or another independently assessed indicator, which can be considered an important criterion (i.e., 'gold standard'). The psychometric literature distinguishes two main types of criterion validity:

- ***Concurrent validity***

When the outcome of interest is in the present, concurrent validity is the appropriate property to determine criterion validity. In this case, the scale is administered at the same time (concurrently) as the external criterion is measured, and each are compared. For example, if a scale aims to assess work stress, criterion validity can be established by (concurrently) comparing the outcome with the measurement of cortisol levels in the saliva of the employee(s), as this is considered the gold standard for a work stress indicator.

- ***Predictive validity***

When the outcome of interest is in the future, predictive validity is the appropriate property to determine criterion validity. Predictive validity reflects the extent to which scale can predict future outcomes. For example, a scale used for employee selection is expected to predict an employee's future performance. In this case, the scale's criterion validity can be established by comparing the outcomes of the assessment in the present with an objective measure of performance in the future. It's important to note that the primary aim of most people assessment tools is to predict future performance or behaviors, making predictive validity one of the most relevant psychometric properties.

Closing Remarks

This guide introduced the core psychometric concepts needed to evaluate the quality of assessment tools in organizational settings. Reliable and valid measurement is essential for evidence-based decision-making. Because no single statistic captures an assessment tool’s quality, psychometric evaluation requires evidence from multiple forms of reliability and validity. A basic understanding of these concepts helps practitioners avoid poor-quality tools, ask critical questions, and make better evidence-based decisions.