2
Types of psychological tests and their validity, precision and standards

Tests are usually classified in objective or projective, according to Pasquali (2008). In case of projective tests, a person is asked to have a certain behavior and project his/her latent traits on a paper sheet or object – as when describing what a person sees in Rorschach test. Normally, there are no right or wrong answers to projective tests: they are open to any possibility. The results are compared with those of other people and deeper investigations are performed when the answer of a respondent does not agree with most of the answers in a normative sample.

Though not always the case, criticisms of projective tests include some discrepancy between statistical and clinical validity. The criticism of lack of scientific evidence to support them has been referred to as the "projective paradox" (Cordón, 2005). It is usually said that projective tests rely too much on clinical judgment, lack proper statistical reliability and validity and have little standardized criteria to which results may be compared. The fact that projective tests are mainly used in clinical realms influenced some to think that projective tests could be validated in the clinical context itself, with no need of psychometric studies. In this respect it is also said that psychometric evaluation of projective tests could lead to impoverishment and categorization of their content and that would be contrary to the core intention of clinical evaluation (Cordón, 2005). The risk of impoverishment and categorization may have directed some psychologists to the opposite extreme, that is, refusing psychometric evaluation in their research. A dichotomy was created, then, between professionals that criticized the projective techniques of evaluation because they did not make use of systematic statistical methodology and those who fully trusted psychometrics possibilities in any kind of evaluation context.

Pasquali (2008) considers that a false issue. He thinks that it might be more challenging to use instruments of psychological evaluation in the clinical context because of its complex, idiosyncratic and ambiguous character. Projective instruments typical of clinical contexts are harder to quantify and standardize if compared to objective tests. Nevertheless, that would only mean more work would be necessary to perform psychometric studies. Actually, there have been many empirical studies based on projective tests (including the use of standardized norms and samples), particularly in the more established tests. Exner (1993), for example, performed hundreds of validity studies about interpretations of Rorschach test.

Objective tests, diversely, have standard answers and are basically divided in direct observation or self-report tests (Cohen & Swerdlik, 2009). Tests for direct observation are those in which the psychologist asks the respondent to perform a task or behavior and the psychologist is responsible for registering the respondent's score. Self-report tests (or self-report inventories) are those in which the very person responds test items (Cohen & Swerdlik, 2009).

Objective tests fully depend on the concept of precision. It is a very important characteristic of psychological instruments. Precision studies are a systematic way of evaluating error in measure. Since mistake is a possibility for any evaluation, being able to estimate the magnitude of the error is of paramount importance. Precision studies provide a new opportunity of evaluation and are an attempt to guarantee that the attribute tested has not changed between test applications. The objective is verifying fluctuations in test scores under similar application conditions. This way, it can be defined as how much test scores are immune to fluctuations that occur because of unexpected, irrelevant and/or undesirable factors (Pasquali, 2008).

Psychological measuring are always vulnerable to error and the practical goal of precision evaluation is what error magnitude is tolerable so that the measure is not disposable. Several sources of error are possible, *e.g.*, subjectivity in test application, differences in evaluation contexts, problems with the content of the tasks used for testing, and others. Therefore, in Brazil, the Federal Council of Psychology (CFP, 2003) requires specific precision analysis in order to

consider a test valid – equivalence (parallel forms), internal consistency, test-retest reliability, precision of evaluators, besides inquiring if the coefficients derived from such procedures are calculated for difference groups of subjects (CFP, 2003).

Though precision is necessary, it is not sufficient for the validation of an instrument. Tests with low precision may be influenced by many sources of error and that makes it hard to identify if score fluctuations are due to important or irrelevant factors. This way, scores are not very reliable and compromise the validity of the test interpretation. On the other hand, even though high precision means little vulnerability to error sources, it is not sufficient evidence that the interpretations associated to the scores are legitimate. High precision is, therefore, only the first step. Validity analysis is necessary to prove that the test really is evaluating whatever latent trait it was supposed to (Pasquali, 2008).

Validity is a fundamental characteristic of psychological tests. It attests whether interpretation made upon data collected through a test is legitimate, *i.e.*, if there are clear data to indicate that a certain interpretation is accurate and result from research planned specifically to test the assumptions of such interpretation. Validity refers to the scientific basis of psychological instruments. Therefore it justifies the relationship proposed between indicators and psychological characteristics (Muniz, 2004).

There are several ways to study the validity of test interpretation. They may be based on the test content – content validity – and refer to the extent to which a measure represents all facets of a given construct. For example, a depression scale may lack content validity if it only assesses the affective dimension of depression but fails to take into account its behavioral dimension. Consultation of experts in the area is fundamental for the decision of whether the content of a given test is fully valid.

Tests are also validated with regards to their constructs. Construct validity is the degree to which a test measures what it claims to be measuring. Constructs are abstractions created by researchers in order to conceptualize the latent variable, which, though not directly observable, is the cause of scores on a given measure. Construct validity examines whether the measure behaves like the theory says a

measure of that construct should behave. For that, several procedures may be adopted – convergent-discriminant validity (correlation with other tests), differences among groups, multitrait-multimethod matrix (MTMM), internal consistency or factor analysis (exploratory or confirmatory) and experimental design (Primi, 2003).

Criterium validity is the last aspect according to which psychological tests are studied and refer to the extent to which measures of a test are demonstrably related to concrete criteria in the "real" world. This type of validity is often divided into 'concurrent' and 'predictive' sub-types of validity. The term concurrent validity is reserved for demonstrations relating a measure to other concrete criteria assessed simultaneously while predictive validity refers to the degree to which any measure can predict future. In objective tests validation must be predictive.

Three categories of psychological tests are then known: (1) projective tests, (2) self-report and objective tests and (3) objective tests with direct observation. Although not necessarily psychological, screening is an important part in the realm of instruments studied by psychometrics. Some screening instruments aim at evaluating behaviors directly. Screening tests, however, have different characteristics other than the above-mentioned projective tests, self-report and objective tests. An example is measuring observations of other people, such as the Behavior Assessment System for Children - BASC (Reynolds & Kamphaus, 2011). BASC has three scales: (1) assessment by parents, (2) assessment by teachers, (3) self-report tests. The first two measures are, by definition, objective tests of assessment of others, *i.e.*, indirect observation. Those are not part of the list of psychological tests by Pasquali (2008) or Cohen and Swerdlik (2009). However, whether or not being a "real" psychological test, measures such as BASC must undergo rigorous psychometric analysis to be considered ready to assist professionals in intervention decisions.

Though precision is necessary, it is not sufficient for the validation of an test. Tests with low precision may be influenced by many sources of error and that makes it hard to identify if score fluctuations are due to important or irrelevant factors. This way, scores are not very reliable and compromise the validity of the

test interpretation. On the other hand, even though high precision means little vulnerability to error sources, it is not sufficient evidence that the interpretations associated to the scores are legitimate. High precision is, therefore, only the first step. Validity analysis is necessary to prove that the test really is evaluating whatever latent trait it was supposed to (Pasquali, 2008).

Psychological measuring will always be vulnerable to error and the practical goal of precision evaluation is what error magnitude is tolerable so that the measure is not disposable. Several sources of error are possible, *e.g.*, subjectivity in test application, differences in evaluation contexts, problems with the content of the tasks used for testing, and others. Therefore, in Brazil, the Federal Council of Psychology (CFP, 2003) requires specific precision analysis in order to consider a test valid – equivalence (parallel forms), internal consistency, test-retest reliability, precision of evaluators, besides inquiring if the coefficients derived from such procedures are calculated for difference groups of subjects (CFP, 2003).

Another important aspect of psychological tests is standardizing the interpretation system of test scores. Results from other tests – similar ones – become reference groups and are used as standards against which the results of the new test are compared. This way, it is possible to define what results are very likely or unexpected (Pasquali, 2008). For example, Beck's Depression Inventory – BDI – (Beck, 1998) evaluates depression against normal behavior, that is, the average score of individuals who do not have depression. The results of the reference group are the standards for the comparison of the results of a tested person. Indeed, test scores are usually compared to the scores of normative groups (the latter case) and also to results of groups who are expected to present the researched latent trait. In BDI, for example, it is possible to compare the results of a tested person with the results of groups with depression and groups without depression. The professional is able, then, to decide which group results are closest to the condition of the tested person.