What constitutes a good test?

First be clear on your own objectives

Spend considerable time on identifying and describing which knowledge, skills, abilities or other characteristics you want to assess. Use these as a firm guideline against which you set out to find a test that can closely match your needs. Use professional help to assist you in sifting through thousands of tests that are available in the market.

Be wary of a vendor who describes their test as measuring many seemingly unrelated factors. Also be wary of tests that consist of a limited number of items yet several attributes, for example a 30-item test that purports to measure 10 different factors – you're looking for at least seven items per test scale; double that is good going.

Once you have narrowed down your search to a handful of tests, you are ready to put each to the test (excuse the pun), using eight handy pointers.

1. Inspect the thoroughness of the test manual – the spine of assessment

A good psychometric test is accompanied by a user's manual that describes the important characteristics and qualities of that test, and a technical manual in which its reliability and validity are detailed and what factors to consider when using the test. The purpose of the manuals is to assist and guide users in planning and reviewing years of development work done by professionals, thus be sure to familiarise yourself with the manuals before administering a test, and continually use them as a reference, no matter how experienced you become in a particular test.

Manuals should describe which materials are required for administration (e.g., test booklets, answer sheets, scoring keys, etc.). Administration instructions should include issues such as the strict control of testing conditions (e.g., noise, lighting, time) and how to handle administrations that risk any deviation from standard procedure.

Qualifications for administering should also be clearly stated in the manual. Some tests have a time limit, while others allow the test taker unlimited time to complete the test. The purpose of time limits and how they were determined should be documented. For tests that have time limits, greater administrator training may be needed. Think about what your scoring and test reporting needs are, be it on-site, or self scoring (either by hand or by machine), or whether you are required to either call in, mail, or fax the test results to the test publisher for scoring. Consider which test-reporting options are most suitable to your needs.

Be sure that someone on your team is qualified to administer and interpret the results prior to purchasing the test. Some tests require the test administrator or individuals interpreting test scores to have certain academic credentials (e.g., MA, PhD) that reflect coursework in statistics, test interpretation, or test development and validation.

These and other practicalities should all be detailed in the test manual.

2. Consider how the test was developed

What was the test author's experience that led to the creation of the test? What is the theory on which the test is based? Consider whether these experiences and theoretical framework match with that of your own. Was the test developed on people that are similar to the employees or applicants of your organisation? These people will likely form the norm group against which the scores of your test takers will be compared. What process was followed during the development of the test?

Answers to these questions provide information on the logic, care, and thoroughness by which the test was developed. References that back the educational background and work experience of the persons who developed the test should be documented. This facilitates confidence in the test and becomes vital in the event of a legal challenge.

3. Look for firm evidence of reliability

Reliability refers to the consistency or stability of test results, in other words, whether test results will be highly similar when administered over time or in comparable but non-identical situations. Test results are reliable to the extent that they are free from random error. For example, if you were to measure a person's IQ as 115 at a particular time and 107 at a later time, and then 115 after a third assessment, the test would not be very reliable. If, however, in a series of measures, you got the same measure (say, 110), the test could be described as reliable – even if it may not be accurate (i.e., valid) and the person's IQ is really 120.

There are several ways to assess the reliability of a test example. For more information, please read *Testing and Assessment: An Employer's Guide to Good Practices* posted by O*Net (www.onetcenter.org).

4. Look for statistical proof of (especially) criterion-related validity

Validity refers to the accuracy of the inferences made based on test results (e.g., how accurate it is to say that a higher test score indicates that a person is more likely to be a better performer). A valid test measures what it is supposed to measure, and nothing else. Test results are valid to the extent that they are free from systematic error. For example, say we want to measure people's emotional intelligence (EQ). If all we had was a test that measures personality style, we would assess the people and record the results. Even if the measures were highly reliable, that is, consistent from one measure to the next, they would not be valid.

The measures wouldn't be completely useless, however, because there generally is some correlation between EQ and personality attributes. Although we sometimes have to try to get by with proxy measures, there is no doubt that an EQ measure would be more valid for measuring emotional intelligence than a personality measure.

A test should adhere to various types of validity. Content-oriented validity may be demonstrated through a statistical link between the attributes of the test and the requirements of the job. Construct-oriented validity details how the test relates to other tools measuring the same notion. A test manual should provide ample evidence of criterion-related validity. For example, research may show a statistical relationship between test scores and some outcome of interest (e.g., supervisory ratings of job performance, average monthly sales, turnover).

Further validity information is outlined in *The Standards for Educational and Psychological Testing* (www.apa.org/science/standards.html). In addition, check out the *Principles for the Validation and Use of Personnel Selection Procedures* (www.siop.org). Trained professionals can help interpret whether sufficient scientific evidence exist regarding a particular test to support your assessment objectives.

5. Think through the implications for potential test bias and fairness

The test should be related to outcomes in a similar manner for all individuals, regardless of gender, race, or other classifications. This does not necessarily mean that the test must have similar results for different groups of people, but rather that it is not a biased indicator of an outcome of interest. For example, in general, more women than men will score high on a test scale that measures caring and nurturing qualities. The test, however, would not be considered biased if women and men with similar scores achieved similar performance on that aspect of the job.

Test results should be interpreted and applied fairly to ensure that discrimination of any kind and unfair labour practices do not occur. Professionals involved in people management and decisions should establish clear criteria for assessment prior to establishing the assessment procedure and tools to be used. For example, a thorough job analysis will reveal the critical and desired competencies needed against which a test can be selected and assessment can take place.

6. Ponder what baselines (norms) exist for interpreting test scores

A test score *per se* is virtually meaningless. For example, a derived score of 100 may mean average performance on an IQ or EQ test, maximum achievement on an ability test that displays performance in percentages, and a non-preferred style on a personality component of a test.

Whether a test score is considered strong or a challenge depends on the distribution of scores of a comparison group. This comparison group is typically referred to as a norm group, and forms the benchmark against which individual test scores are interpreted. This process of benchmarking is called standardisation.

Test publishers should provide information about the different norm groups that are available for the test being considered. General norm groups should be large in size (including thousands of people's test results) and demonstrate a fair representation of broad population characteristics. Specialised or custom norm groups may be somewhat smaller in size, but still be representative of the target population. Search for and use a norm group that is similar to the group of people that are in the position for which testing is being used. When comparing the test results of two or more individuals with one another, always make sure that the same norm set is applied to standardise the test scores.

There are other ways to interpret test results, including expectancy charts and cut scores. These are developed based on information about how the score relates to outcomes of interest. Test publishers should include guidelines in the manual on data that can aid in appropriate test score interpretation.

7. Is the test up-to-date?

How many years ago was the test developed? When last was the test updated? A good test is often a life's dedication and hard work, and is never fully completed. The best tests have decades of research years behind them and several updates as new insights are gained. Test publishers should indicate when the test was developed and when the test was last updated. Vendors often update their tests to comply with new legal requirements or to reflect changes in vocabulary or terminology.

Be wary of tests that have a multitude of spin-offs that are far away from the original purpose of development. For example, a test that was originally developed to assess an individual's career interest, but now has spawned a test that assesses a person's leadership potential based on the same test attributes, may be too far a stretch to have the same rigorous psychometric properties. The new test cannot claim reliability and validity on the basis of that of the original test, as the purpose and circumstances under which it is used, have changed substantially.

8. Value for money

Direct costs of the test usually include test booklets, answer sheets and a test administrator’s manual. Should the test be hand-scored, a scoring template will be reflected in the cost as well. Computer-based testing generally includes licensed software, and a prepaid number of uses. Cost for test scoring often includes the generation of a report – make sure the type of report is the one that you have selected.

Shop around for different price structures to determine what the best value for money will be. A more expensive test or test report is not necessarily a better one; some reports are heavily padded with the same information that are presented in different formats. Check out the level of detail that is provided in a report: are the interpretations very generic, or do they rely on complex algorithms that are well thought through?

Testing fees usually need to be considered as an ongoing expense that should form part of one's annual budget planning. Few test publishers will license a test for unlimited usage. Part of the cost of the test is for the substantial investment that test publishers made in researching and developing a high quality measure. Creating a high-quality and effective test requires time, money, and people to research and develop, revise, validate and update.

Proper use of the test

We share the responsibility to use a test only within the purpose domain that it was developed for and to adhere to the guidelines as set out by its test publisher. Good tests start with their author; their continued value depends on the professional accountability of those who use it.