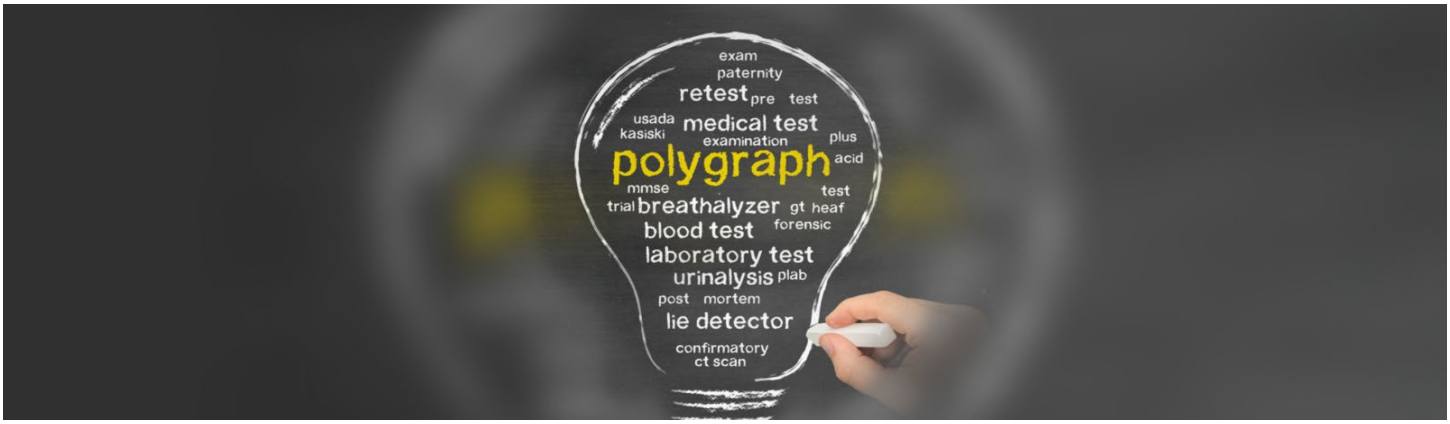

Five Minute Science Lesson: What is a Standardized Test?

Raymond Nelson

A test is a procedure used to characterize, quantify, identify or classify a phenomena of interest. Tests are also used to make predictions – about a future outcome or past cause related to the available data – which can be thought of as a special form of classification. Tests are often used when direct physical measurement – which requires both a physical phenomena and defined unit of measurement – is not possible. Scientific tests often make use of proxy information for which there is a theoretical or statistical relationship with the phenomena of interest. For this reason, scientific tests are often inherently probabilistic – including when test results are reduced to categorical results – and are not expected to be deterministic or infallible.

A desirable characteristic of a scientific test is good validity. That is: the results pertain to reality in some useful way. But a test cannot be valid if it is not first reliable. That is: test results are consistent when the test is repeated or test data is re-evaluated. Some tests such as academic achievement tests used in schools are intended to be sensitive to change. Some psychological tests are also intended to be sensitive to change such as tests for mood or affect disorders. Other psychological tests such as those for personality and intellectual functioning may be intended to be quantify characteristics that are regarded as stable. A test can be reliable (i.e., consistent) but not valid, but a test cannot be valid if it is not reliable. For this reason, it is said that the reliability of a test sets the upper limit of its validity. Reproducibility of test results is often directly related to the degree of objectivity, or latent subjectivity, in the analytic process. Standardization is an important strategy for achieving test reliability.

A standardized test is a scientific test – whether used in psychology, education, industry and government – that is administered to each individual in the

exact same way under the exact same circumstances. In other words, each person is asked the same questions in the same manner. Standardized tests are used to measure everything from intellectual functioning, aptitude, achievement, personality, interests social characteristics, physical abilities, and social behaviors. Another characteristic of a standardized test is that test data are evaluated – scored and interpreted – using consistent methodology for which the assumptions and procedures have been shown to produce outcomes with a satisfactory degree of validity.

Standardized tests are used today in psychology, education, industry and government. The first use of standardized tests was in China during the Han Dynasty (206 BC to 220 AD) with the goal of selecting people for civil service roles based on merit instead of family connection. Standardized tests have been used in U.S. schools since the later part of the 19th century, often to determine readiness or admission to a college or university. Standardized tests for intelligence were first developed in the early 20th century – as a replacement for earlier, erroneous, attempts to measure intellectual ability as function of the physical size of the human skull – in order to identify school children that are lagging behind others and may benefit from extra assistance. Intelligence testing of adults became more common with the entry of the U.S. into WWI in 1917, when a group of psychologists helped to develop the Army Alpha and Army Beta tests, for literate and non-literate recruits, which would be administered to over 1.7 million persons in the next year.

Data for individual test results can also be compared to a criterion standard. For example: can a fire-fighter applicant can drag a 165 pound rescue manikin 100 feet in less than 60 seconds? Or: can a fire-fighter applicant complete a timed obstacle course, requiring a combination of strength, agility, and endurance? Criterion data can then be aggregated for groups, allowing for statistical comparison. What proportion of fire-fighter applicants can complete the strength, agility and endurance tasks successfully? Data for any individual can also be compared to statistical norms. For example: what proportion of fire-fighter applicants require more time or less time to complete an obstacle course?

Statistical norms can be developed as a function of the underlying theory of a test. For example: a university can test for institutional gender bias by comparing the number of male and female admissions with the number of male and female applications using a binomial test of proportions and comparing the result to a statistical reference distribution for unbiased outcomes. The probabilistic result can be thought of as either a density – the probability of obtaining the observed test data under an unbiased (null) model – or as the cumulative probability of likelihoods larger or smaller than that of the observed data. Data from standardized tests can also be used to investigate hypotheses and to develop theories.

Statistical reference data can also be developed empirically by aggregating the test scores of numerous sample groups using the central limit theorem (CLT). The CLT states that the distribution of sample distributions will be approximately normally distributed regardless of whether the population is skewed. Another aspect of the CLT is that if the sample groups are randomly selected the resulting distribution of sample means will converge towards the unknown population mean. In this way we can calculate reproducible statistical parameter estimates for large populations that cannot be easily measured. For example, what is the average height of American adults? Data for any individual can then be compared with the empirical reference distribution to answer questions such as the proportion of persons that are expected to achieve a certain score, or the proportion of persons that are expected to achieve a greater or lesser score than a single individual score.

Standardized tests are not the only form of test. ***Unstandardized*** tests are those for which the test administration may vary from one individual to another, and those for which the analysis and interpretation of test data may vary from one person to another. Unstandardized tests are characterized by an absence of objective criterion and statistical reference data as a basis for a reproducible analysis and interpretation. Results from unstandardized tests may be subject to criticism for their ambiguity or subjectivity, and for their meaningless or unknown relationship to practical outcomes of interest.

In the field of professional psychology some tests are described as *projective* tests, because they are designed to create a context, through the use of ambiguous stimuli and ambiguous analytic procedures, for a test administrator to project their own expert intuition – which may be subject to a variety of forms of bias – into the analysis and result. Examples of projective tests are the Rorschach inkblot test, Thematic Apperception Test, Draw a Person and Sentence Completion tests. Projective tests are regarded as unreliable, and for this reason have questionable validity. Because they may elicit nuanced information that may not be captured by standardized psychological measures – which often rely on TRUE/FALSE items and Likert scales – projective tests are still considered to be a useful aspect of a complete psychological evaluation. Projective tests should not be used alone and, when used are integrated into a comprehensive psychological evaluation that integrates the results from objective test measures, interview data, and an individual's social/developmental history into a diagnosis, narrative and treatment plan that attempts to address a referral question.

Another alternative to a standardized test is referred to as a *clinical assessment* which refers to the use of unstructured expert judgment based on observation. Clinical assessment, like *holistic assessment*, is a process of synthesizing and integrating various sources of informa-

tion for which no structured procedure is defined. Clinical assessment will have inherent vulnerability to bias along with inherent limitations around reliability (and by extension inherent limitations around validity). Scientific comparisons of effect sizes for unstructured expert judgment have sometimes been found to be no different than judgments from non-experts, and are consistently outperformed by structured and standardized solutions.

Developers of standardized test attempt to eliminate sources of bias that may result from differences or inconsistencies in test administration and test data analysis. To the degree that they rely on proxy information, and do not actually measure what they purport to measure, there is always the potential that unintended factors may influence a test result. When these other factors can be exploited it may be an opportunity for faking. There is no test in existence that will be completely immune to potential faking, and for this reason many tests incorporate features, both data acquisition and analysis, to evaluate for the likelihood of faking. Some tests are loaded for verbal information such that evaluation of other skills may become difficult if verbal communication is difficult. Similarly, tests that rely on cultural information may be unfair for persons who are immigrants. In the most extreme cases, standardized tests in education have sometimes been criticized as testing a person's ability to take standardized tests.

Standardized tests are characterized in part by the publication of standards for test administration and test data analysis. Another characteristic of standardized tests is standardization of the required knowledge, training, skills and experience of the test administrator. Standardization of professional competence – through rigorous education, training and certification – has been, for most of the last century, among the most important strategies for increasing the reliability and validity of a test. Some standardized tests may be administered by a proctor or invigilator, while data analysis is performed by an expert or psychometrician.

Today, with powerful and inexpensive microcomputers available virtually everywhere, inconsistencies in test administration and test data analysis can be largely reduced as a factor in test reliability through the use of *computer assisted testing* (CAT) designs in which test stimuli, and even instructions can often be automated. CAT is also widely used to reduce human error and subjectivity in test data analysis. In the future – with the increased availability of machine-learning/artificial-intelligence algorithms that can rapidly develop and test hypotheses without little or no operator input – we are also likely to observe the increased use of *computer adaptive testing*. For example: a testing algorithm may select subsequent test stimulus questions based on past responses that were successful or unsuccessful, adjusting the content so that the likelihood of success equals or exceeds a 50% level (or some other level as suits the context and testing objective).

The idea of standardized testing is that it gives give everyone the same chance to produce a satisfactory outcome – with the practical result that, with the assumption that an individual is reasonably represented in the test development, statistical reference data, and criterion standards, we are permitted to assume that observed differences in individual outcomes are a reflection of individual differences. Regardless of the intended objectivity of standardized tests, the potential for unknown bias is such that there will always be ongoing discussion about the degree to which tests can achieve the egalitarian ideal of unbiased accuracy.

As we move into an era in which there is increased interaction of humans and intelligent machines, there will also be important discussion about the degree to which human professionals, whether test administrators or machine/algorithm developers, are regarded as responsible for the professional conclusions that may be based in part on standardized test results. Regardless of whether administered by human experts, technician-operators, or automated systems, standardized tests – including standardized test administration and standardized test data analysis methods – ensure that outcomes are reproducible and reliable, leading to an ability to achieve a more satisfactory understanding of the criterion validity and practical value of observed test results.