Classical Test Theory (CTT)

Gerianne de Klerk

Introduction

Today most people will have seen or even completed psychometric tests in their lives. These could be for instance 'just-for-fun' tests you can sometimes find in magazines, school tests or recruitment and selection tests. Psychometric tests are widely available and measure a large variety of constructs (e.g. intelligence, personality, motivation). But how can you know whether a test is actually a good test? How do you know if a test measures a construct accurately? Or whether that test actually measures what it says it measures? And what about influences from outside the testing situation, what kind of effect will they have on the outcome of a test? For instance, one group of candidates is packed into a small room where next doors construction work noise is overwhelming, compared to a group of candidates who complete the same test each in a separate, noise-free room. Do you think these two sets of candidates will have comparable scores, or could the testing situation have an influence on their obtained score?

And then we can think of several scenarios where a test ends up measuring a construct different from what it had aimed for. For instance when only one question is used to measure a complex personality construct, do you think only one question could cover the whole construct and will measure that construct accurately? Or where a test is supposed to measure numerical reasoning ability, but as most of the questions are phrased in a very complex English language, it might actually measure a candidate's ability to understand that form of English, rather than the numerical reasoning ability in question. Classical Test Theory is involved in mapping the relative input of unpredictable influences, during the administration of a test, onto the test result, as well as mapping the systematic characteristics during a test administration of people and test situations.

Classical Test Theory

Classical Test Theory, commonly abbreviated with CTT, originates from the beginning of the 20th century (Spearman, 1910). The final 'Classical Model' was only published in the late 1960s however. The most important formula of Classical Test Theory goes as follows:

$$X = T + E$$

X = Total score/raw score obtained

T = True score

E = Error component

Classical Test Theory assumes that each test score (X) contains a True component (T) and an Error component (E). When measuring a psychological construct, unsystematic errors occur. These unsystematic errors could be anything, for instance distractions from outside the testing situation, physical wellbeing of the candidate or luck/bad luck. You can think of many different influences on how a candidate could be affected in how well he/she is doing at the specific moment of administering the test. Sometimes these influences have a positive effect on the test result; other times they have a negative influence. In other words they cause a band of error around the True score.

The True score can be seen as the systematic component of the raw score obtained. Classical Test Theory assumes that the errors in measurement are divided equally around the average, i.e. the deviations are to both sides of the True score equally high. The True score then is the average score. This means the average error of measurement is 0, as the coincidental positive and negative deviations average each other out.

Systematic errors are different to unsystematic errors. With systematic errors we think of a characteristic of the test or the testing situation that will affect all measurements equally. You can think of a mistake in one of the test items which will be presented to all the candidates completing the test, therefore influencing all candidates the same way. As psychological tests are mainly used to determine individual differences, the influence of systematic errors is unimportant and is not included in the Classical Test Theory concepts introduced hereafter. However it is good to note that when candidates using a test with some systematic error are compared with candidates who completed a test free from such error the comparison will be unfair. Later on I will briefly come back to the subject of fairness in testing. Let us first go back to the concept and consequence of unsystematic errors from Classical Test Theory.

In the hypothetical case that we administer one test repeatedly to one candidate in an almost indefinite number of times, the range of measurement error is equal to the range of observed scores. In other words, the candidate will have an average score over all these repeated measures and the difference of his/her lowest score obtained with this average indicates the largest negative error. In this case the error represents the situation where the candidate had the most negative external influences which caused him/her to perform specifically poorly compared to the other administrations. This also works the other way around in having positive external influences on the performance of the test.

Unfortunately, in real life it is impossible to have such a situation with repeated measures. With most, if not all psychological constructs, learning and memory processes are involved which will have a systematic, but unwanted influence on repeated measures of a test. For instance, people can remember their previous test session and answer similarly, or they will learn how to solve certain problems and perform better on the test the next time (e.g. on ability tests). This is different from other types of measurements, for instance a measurement of length or time. Such measurements are constant and have a very small error of measurement. A psychological measurement involves much more complex psychological processes, which are more difficult to describe and measure in a test as accurately as physical processes could be.

However, errors of measurement will also average out over a large number of repeated measures, when not one but a large group of people is used, each individual having completed the test once. This error of measurement should not correlate with characteristics from within the population of people however.

From the above formula and several of these assumptions, formulas for reliability and standard error of measurement can be derived. They are central to Classical Test Theory and with these two concepts an estimate of the accuracy of a measurement can be obtained. It is important to realise however that a test with a high level of accuracy (or reliability) does not necessarily measure what a test is supposed to measure. I go back to the example used in the introduction; the numerical test using very complex English language might actually measure a candidate’s ability in that language rather than numerical reasoning ability. However, it might measure this English language ability quite well and accurately, therefore the reliability of the test could be high, but the construct it measures is not the construct it is supposed to measure. This issue will be explored in the Validity section. First we will focus on reliability.

Reliability of a test

The reliability of a test, or how accurate a test measures, can be estimated in two ways:

1. Through repeated measures
   a. Parallel form method (two tests, which are parallel forms)
   b. Test-retest method (same test)
2. Through single measures:
   a. Split half method (two halves of a test)
   b. Internal consistency method (item by item)

The parallel (or alternate) form method assumes complete equivalence of two tests; you could actually exchange the tests and candidates will score both tests identically. The correlation between the total scores of these tests form the reliability of the independent test scores. This method is however very complicated because items in the two tests should be equal but they cannot be identical (otherwise you will have exactly the same test). You can prove if tests are parallel forms of each other if their average observed scores are identical, as well as their variance in scores. Also, both tests should have equal correlations with any other variable.

To estimate test-retest reliability, one test will be completed by the same group of people twice. However, there will be a considerable time gap between the two administrations. The correlation between the two total scores is an estimate of the tests reliability, but only if the two administrations can be seen as independent of each other. This method can give a good estimate of reliability, but is not always a good indication if there could have been changes in the measured characteristic since the first administration of the test. For instance, there could be a learning effect. Candidates might have looked up or discussed some of the items and learned how to solve it by the time the second administration takes place. This will decrease the correlation, and therefore the reliability estimate of the test. Also candidates might feel the need to be consistent in their answering pattern, with that making the two administrations unnaturally consistent and inflating the reliability estimate. Another issue with this approach is determining the time gap. A too short time interval between the two administrations will come with a bigger chance of people remembering the previous administration, and subsequently learning or memory effects could play a part. Too long a time interval might cause the sample group to decrease in size as some people might not be interested in taking another test or they cannot be found anymore.

The split-half method relies upon one test and is more efficient than the alternate form method. The test is split in two equal halves that are equal in length and preferably equal in difficulty (parallel halves). For each half a total score is calculated and when both halves are truly parallel, the correlation between the two halves is an estimation of each half test's reliability. To determine the full test's reliability, a correction needs to be applied to the half test's reliability. This correction is dependent on the full test's length and how that relates to a reliability estimate. This can be done using the Spearman-Brown formula which relates reliability to test length (you can look up the formula in Wikipedia: http://en.wikipedia.org/wiki/Spearman-Brown_prediction_formula). If the test halves are not completely parallel, the estimated reliability figure obtained through this method will be an underestimation of the true reliability of the test. To prevent problems arising out of two halves being not completely parallel, the reliability can be estimated using the internal consistency estimation method.

The internal consistency method is based on the notion that individual items in a test can be swapped around. The test is administered once to a representative group of people. Subsequently covariances are calculated between all the items as well as the variance of the total score (a covariance is a measure of the relationship between two variables or how they vary together. A positive value will be obtained when both variables have a value deviating from their mean in the same direction (if one is high, the other is high too)). The most common internal consistency measure is the alpha coefficient, also called Cronbach's Alpha (Cronbach, 1951)). It is based on inter-item correlations. The internal consistency will increase if the number of similar items increases as their inter-item correlations increase (i.e. when they correlate they measure a common core). In other words, the longer a test containing similar items, the higher the internal consistency - reliability estimate. For this reason it is always important to ask yourself if the reliability estimate is high due to many, low inter-correlating, items, or if it is because there are a few, highly inter-correlating, items. The second option, a test with fewer, highly inter-correlating items is in most cases preferential.

Hearing all these forms of reliability estimates and knowing how to estimate them, the question springs to mind what a good level of reliability actually is. When can you be confident a test is an accurate measure? There is no agreement on the level that should be obtained for a good test, but some guidelines can be given. First of all you have to think about the purpose of the test. Is it going to be used for high-stakes selection decisions (important 'hire' or 'not-hire' decisions are made in this situation)? You can think of intelligence tests here used for selection procedures where there is a pass/fail decision to be made. Or is the test more of a descriptive nature, like personality questionnaires. A very high reliability figure indicates that the test has very similar items; it might even be quite a narrow band of items that all measure one very narrowly defined construct. Each item does not give a lot of new information on the construct in this case as they all measure the same thing. A very low reliability figure indicates that the items are quite varied; the items are different from each other or might even be ambiguous. In general, values above 0.7 are seen as acceptable reliability figures, whereby for high-stakes testing you would like to see the tests reliability above 0.8. However, be critical of reliability values above 0.9: ask yourself if the tests items are too narrowly defined. For personality type questionnaires, where the constructs are more broadly defined, a value below 0.8 (and occasionally even below 0.7) is acceptable. A reliability figure below 0.6 is generally seen as too low.

Validity of a test

The validity of a test can also be described as the extent to which the test measures what it is supposed to measure. In other words, is the test measuring the construct that you want it to measure (i.e. verbal reasoning ability) or might it be measuring something else than this verbal reasoning ability? Two main goals can be identified in validation research. The first goal has to do with researching the theoretical construct itself. The second goal is about the prediction of behaviour or achievements outside the test achievements. In other words we would like to make a prediction about facts of which we do not have direct evidence, but we can make a conclusion about it as we have knowledge of other data (achievement on the test). The most common forms of validity are:

- Face validity
- Construct validity
    - Criterion validity:
        - Concurrent validity
        - Predictive validity
    - Convergent validity
    - Discriminant validity

Face validity is a very simple form of validity; it researches if the measure, on the face of it, seems to measure what it is supposed to measure. It is a subjective impression of the tests content by the psychologist or even an outsider. It will give a good indication if the test's items seem to be linked to the construct it should measure, but other forms of validity are required to give evidence if this is truly the case.

Construct validity is a more thorough form of researching whether the test is measuring the construct properly. It involves quite a complex process and all other forms of validity methods mentioned above can be seen as contributing to construct validity. It is important for the researcher to identify all the concepts that might be an explanation of the achievements on the test. Then to extract stable hypotheses from the theory, which subsequently needs to be tested through empirical research.

In concurrent validity the measure is compared to other measures of the same construct that are available. For instance the test results are compared to supervisor ratings of each candidate on the same construct.

Predictive validity is similar to concurrent validity, but it involves to what extent the achievement predictions, made based on the test results, can be proven through observations or data at a later stage. This could for instance be a question like, how well did this test predict success in a job? You can then look at those people who did the test during their recruitment, two years later in their job and how well they are performing. It asks if the test predicts a candidate's performance in certain abilities.

With convergent validity an indication of the tests ability to measure the construct is given by establishing whether the test returns results similar to another test measuring the same construct. Obviously, to give a proper indication of this type of validity, the test against which it is compared should have good validity and reliability figures.

The final form of validity discussed here is called discriminant validity. With this form of validity you reason from a different perspective. You are not researching whether the test is measuring what it is supposed to measure but you are researching whether the test is not accidentally measuring what it should not measure. For instance the example used in the introduction about the numerical reasoning test which might measure ability in the English language rather than numerical reasoning ability. If the test has a low correlation with established verbal reasoning or verbal comprehension tests, you have proof that it at least is not measuring those unwanted constructs that you thought it might measure.

Validation research very often show quite disappointing results. Correlations for different criteria are on average not a lot higher than 0.30 to 0.45. Why is that? First of all it is essential that the measures used, both the test to be validated as well as the criterion (other measure or evidence of the construct) have a good reliability. When the test is not reliable, you will not find good validity figures. It can also work the other way around though; a perfectly reliable test can have very poor validity, as argued in the Classical Test Theory section above but in this case you have to search for another explanation of poor validity figures than unreliability of the test. Regarding criterion measures, especially in concurrent or predictive validity studies, supervisor ratings or school grades, which are commonly used, do not tend to be the most reliable. It is relatively straightforward to illustrate the impact that reliability, or rather the lack thereof, has on the validity coefficient. When for instance your test has a reliability figure of 0.80 and your criterion measure is perfectly accurate (reliability of 1.00) the maximum validity figure you can find in your data, i.e. perfect relation, will be 0.89. Be aware of the word maximum here, even when the relationship between test and criterion is perfect, it can never reach a figure of 1.00. When your criterion measure holds a reliability figure of only 0.50, the maximum validity figure obtainable will be 0.63. The lower the reliability of both test and criterion, the lower the maximum validity figure will be.

Secondly, the relation between the test and the criterion could be non-linear. For instance the situation when a low test result goes together with a low criterion achievement, a higher test result with a higher criterion achievement but a very high test result with a lower criterion achievement. As an example of this situation we can use the relationship between motivation and achievement. A low motivation will result in low achievement, a higher motivation in higher achievement but very high motivation can give the candidate stress and tensions resulting in a poorer test result.

Thirdly, there might be a bias in the test and/or criterion measure. For instance the test might correlate well for males but not for females, therefore a heterogeneous population (with both males and females) will not correlate very well. This form of bias is called gender bias. There are many forms of bias. One most commonly found form of bias is cultural bias (see section on cross-cultural testing). The psychological construct might not have the same meanings in other cultures or the items might not be interpreted the same way by people of other cultures, let alone by people who are speaking another native language than the language of the test. Gender or culture could therefore be a moderator on (or suppressing) the correlation between the test and the criterion.

Fourthly, it is important to critically review the criterion measures. Could they possibly be defined too broadly (i.e. when ‘successful performance’ is used as a single criterion measure whilst this consists of several aspects). Or if you would compare performance of a group of candidates in the same position within different companies, is it possible that this position is being fulfilled differently in different organisations?

The validity coefficient found, which will tell you the extent to which the test is valid for making statements about the criterion, can be squared and multiplied by 100 to get to the explained variance in a relationship. To explain this more properly an example: when the predictive validity figure of an ability test is 0.70 with the criterion job retention (i.e. when a person is staying in a job for lets say at least 2 years) it means that 49% ($(0.70)^2*100 = 49\%$) of the differences in job retention, the criterion, is explained (or predicted) by differences in the test achievements.

What to do with test scores?

So now you have developed your test. You have proof that your test has a good reliability and is measuring what it is supposed to measure (validity), but how can you interpret each candidate's score? Can you compare the results of one candidate to the results of another candidate completing the same test? And what about candidates completing different tests, can you compare their results?

First of all, to be able to compare scores of one candidate to the other you have to ensure your test administration is standardised and it is 'fair' to make these comparisons between candidates (this is a requirement even in the trialling and development stage of the test). You want to minimise, as much as possible, external influences that impact on a person's performance on a test. In other words you want to minimise the level of unsystematic error. Every candidate should get the same opportunity to perform at his/her best. So, tests need to be administered under identical conditions. It should not matter where or when, by whom or to whom a test is given, it should be administered in the same way. Therefore instructions in how to complete the test as well as practice items should be exactly the same in each test administration. All candidates should complete the test in a quiet, un-disturbed environment where there are no distractions.

Subsequently, all candidates' performance should be put onto the same scale and their relative performance compared to a representative group of people completing the same test. This group of people, also called the normative sample, should be representative of the target population, i.e. the comparison group should contain the same type of people as the people who are being tested. It is very important to have a large enough comparison group (at least 100 people but preferably a lot more than this) to make sure the sample includes all the typical 'varieties' of people that you would normally find in a population. Be aware that a norm group is not an absolute, unchangeable entity. In time, populations change and regular norm updates should take place for the test.

In most cases, candidate's total scores are compared to the norm (or comparison) group using standardised scores which are based on the comparison group's average and standard deviation. Using the average and standard deviation has the advantage of giving you an indication of the percentage of people who score higher as well as that of people who score lower. The most common example of a standardised score is the z-score. Z-scores indicate how many standard deviations each candidate's total score differs from the average score. If a candidate scores 1 standard deviation above the average score of the comparison group (which represents about the 84th percentile (i.e. 84% score lower)), his/her z-score will have the value of 1. If he/she has the same score as the average score of the comparison group his/her z-score will be 0.

By using z-scores it is easier to compare scores of one candidate on multiple tests and quantify differences between candidates completing the same test. If the candidate scores a z-score of 1 on two different verbal reasoning tests, and if the comparison groups for both tests are the same, you can compare the candidate's score and say something about the relative performance on each test. Comparing raw scores opposed to standardised scores is a lot less informative. What if two candidates scored a raw score of 25 and 28 respectively, what does that tell you? You can say the second candidate performed better on the test, but how much better did he do? If you know the standard deviation of a representative comparison group is 6, than you know the second candidate scored 0.5 z-scores higher than the first candidate. Using the comparison's group average and standard deviation does assume the scores to be normally distributed however. This is a fair assumption to make for most tests but should be tested before standard scores are applied. Once you are satisfied that the distribution of scores is normal, many standardised scores can be used (on top of the z-scores just explained). The next figure will give an overview of some standardised scores in relation to the normal distribution. In here T-scores are mentioned on top of z-scores, which have a mean of 50 and a standard deviation of 10, and standard IQ scores, which have a mean of 100 and a standard deviation of 15.

See Figure on next page

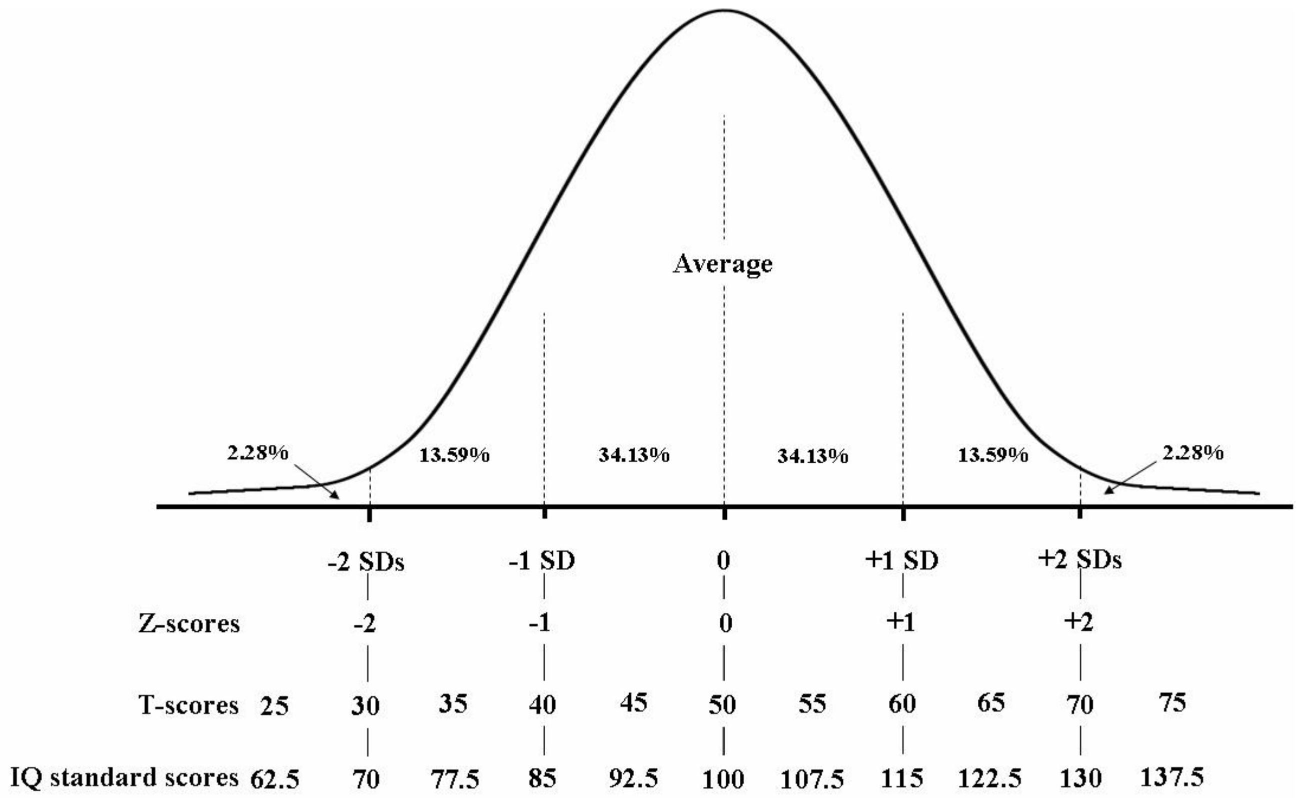

Figure 1: normal curve with standard scores.

In Classical Test Theory, the total score of a candidate is dependent on the content of the test used. It does not give an absolute measure of the characteristic (e.g. ability) of the candidate. A candidate's test score is the sum of the scores received on the items in the test. The difficulty of an item is defined in relation to the comparison group. It is given in the form of a so-called p-value (a number between 0 and 1) which indicates the proportion of the comparison group that answered that item correctly. Candidates have different ability levels and dependent on the test, either an easy or a difficult test, would result in different scores. When a test is difficult, a candidate will appear to have a low ability; compared to his/her performance on a test that is easy, as the candidate will score a lot more items correct and will appear to have a high ability (i.e. higher total score). For this reason it is difficult to compare a candidate's results from different tests unless you have proven that the two tests are completely parallel or equivalent (see reliability section) or using the same norm group and standard scores. This dependency on the comparison group used, determining the difficulty of the test, is removed when using Item Response Theory. This topic is further explored in the Item Response Theory section.