Virtual Test User: Occupational, Ability (Level A)

Course Notes

Contents

Section 1 – Introduction to Psychometric Tests

Introduction

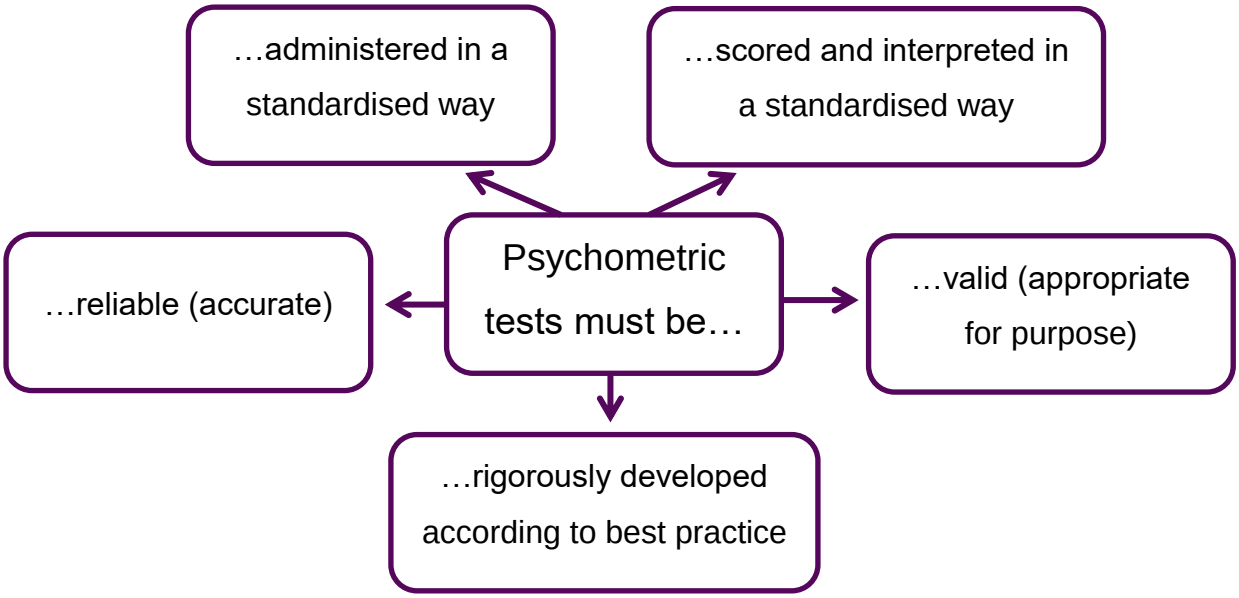
This section will cover the methods of classifying the wide variety of available tests as well as the processes involved in developing and selecting them. This section will also explore the benefits of tests and how to identify appropriate tests for different applications, which include selection, development, coaching, career guidance and team-building.

What is a 'psychometric test'?

Psychometric tests are instruments which have been designed to measure psychological constructs. The term 'psychometric' is derived from two words: 'psycho' relating to the mind and 'metric' meaning measurement and refers to the scientific measurement of psychological constructs.

The British Psychological Society (BPS) defines a psychometric test as:

'An assessment procedure designed to provide objective measures of one or more psychological characteristics.'



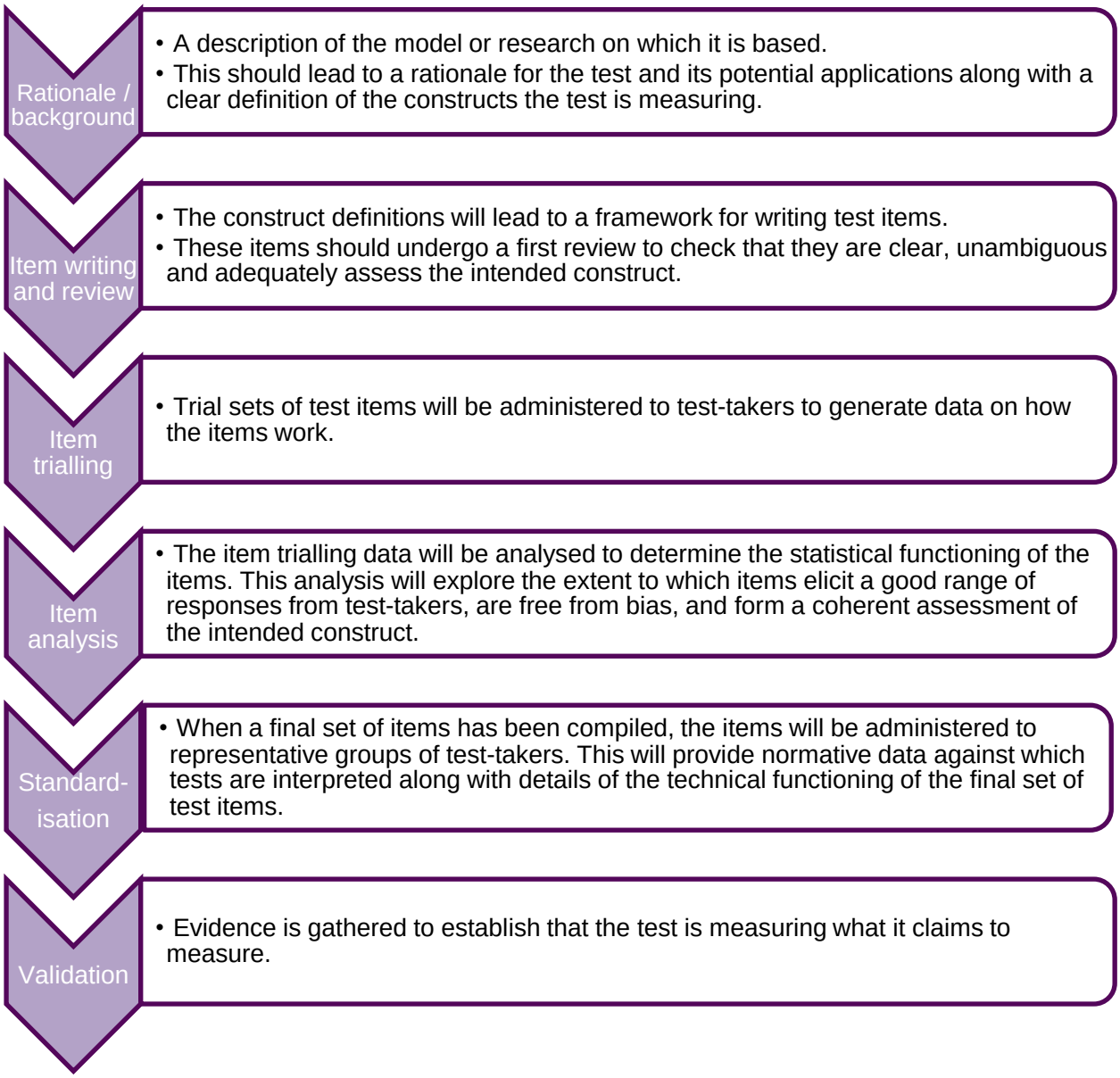
There are many tests and other types of assessment that claim to measure constructs such as ability, knowledge, skill, personality or motivation. Despite assessing psychological constructs, not all of them would be classified as being psychometric as the term only refers to tests which have the properties and meet the technical standards which are outlined above.

This course aims to equip test users with the knowledge to determine whether a test meets these psychometric standards.

The process of test development

In addition to the technical properties which a psychometric test must possess, the tests should have gone through a rigorous development process prior to publication.

The stages of development are:



A test's technical manual or user guide will provide information on a test's development. This documentation is very useful when considering using a new test as it should provide confidence in the test's construction and its properties.

The history of testing

1905

The start of psychometric testing is often traced back to the French psychologist Alfred Binet. The French Government commissioned him to devise a test to measure children's intelligence. He is recognised as the person who produced the first psychometric test in 1905. This was subsequently developed into the Stanford-Binet IQ Test. Binet used the idea of the Intelligence Quotient (IQ) to express test results, ((mental age ÷ chronological age) x 100). Therefore, those with the expected intelligence for their age obtained an IQ score of 100.

1917

High volume ability testing started when Robert Yerkes developed the Army Alpha and Beta tests as a way of assessing recruits for the US Army in 1917. In just a few years, more than a million recruits were assessed with these tests, with the results being used to identify potential in new recruits.

Over the next two decades, research on human ability by Spearman and Thurstone greatly advanced our understanding of the nature of human abilities and their measurement. Many of the fundamentals of psychometrics were established during this time.

1939-45

During World War 2 the use of testing increased with more countries using psychometric testing as part of their recruitment for the armed forces.

In the UK the National Institute for Industrial Psychology (NIIP), promoted the application of psychological principles to industry, including the use of tests for personnel selection. The NIIP also ran training courses in psychometric testing for organisations.

1945-85

Testing in the UK and other countries grew steadily following WW2. In **1946** Cattell developed the 16PF, a turning point in personality assessment. In **1984** the first specific Occupational Personality assessment was published, the OPQ.

1990s

In the 1990s there was a major change in testing. Publishers started to make use of increasingly affordable computers to deliver, score and interpret their tests.

Through the improved efficiency and accuracy of testing, it was the subsequent adaptation of tests for internet delivery that has had the most significant impact on testing. This allows tests to be taken by anyone, anywhere in the world.

Today

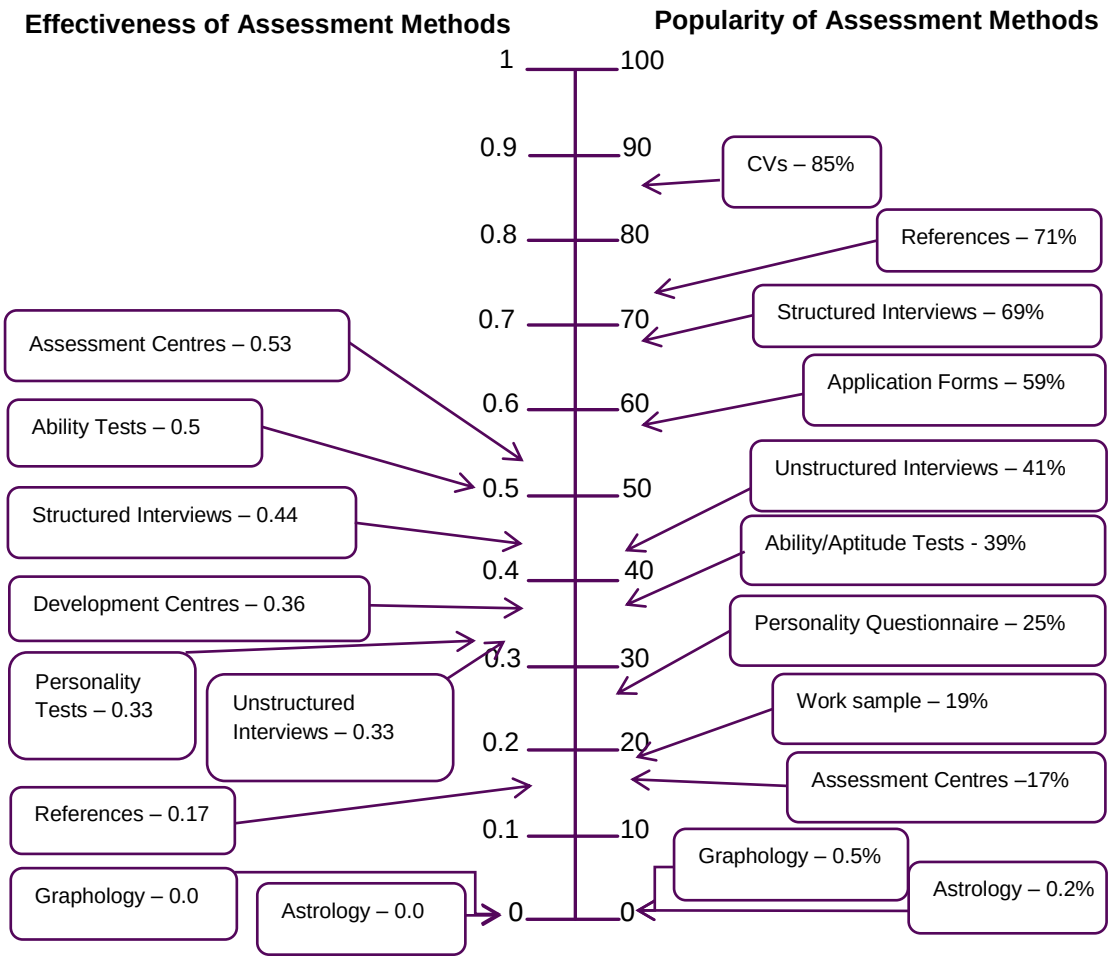
As evidence for the effectiveness of ability and personality testing continues to grow, so does their use. Currently it is estimated that around three-quarters of all organisations use tests for some purposes and this figure is growing.

Why and when should psychometric tests be used?

Tests are used by a large percentage of employers for selection in at least some roles, with 23% using general ability tests and 35% using personality/aptitude/psychometric questionnaires according to a recent survey of 605 UK organisations (CIPD, 2011). These figures do not include the use of tests for development, guidance, coaching and other purposes, so clearly testing is widely accepted. However, interviewing is still more popular than many more effective methods of assessment.

The chart below compares the popularity and effectiveness of widely used assessment methods.
On the left is the effectiveness or 'validity' of the different methods. The figures range from 0 (no value in predicting performance) to 1 (totally accurate predictor of performance), and are based on the research of Schmidt & Hunter, (1998). On the right of this chart are the popularity figures for different selection methods. This information is taken from the research of Zibarras & Woods (2010).

Methods such as interviews and references are widely used, though the effectiveness of references is relatively low and the effectiveness of interviews is greatly enhanced if they are structured rather than being more of an open-ended discussion. Ability tests are one of the most effective predictors, though assessment centres will typically contain a range of measures and sometimes include ability tests. Overall, this chart shows that whilst there is some tendency for more effective methods to be used more frequently, this is not always the case.



Unlike informal methods of assessment (CVs and interviews) which have no clear method underpinning their execution, formal methods like psychometric testing have a solid underpinning method and are objective, accurate and valid.

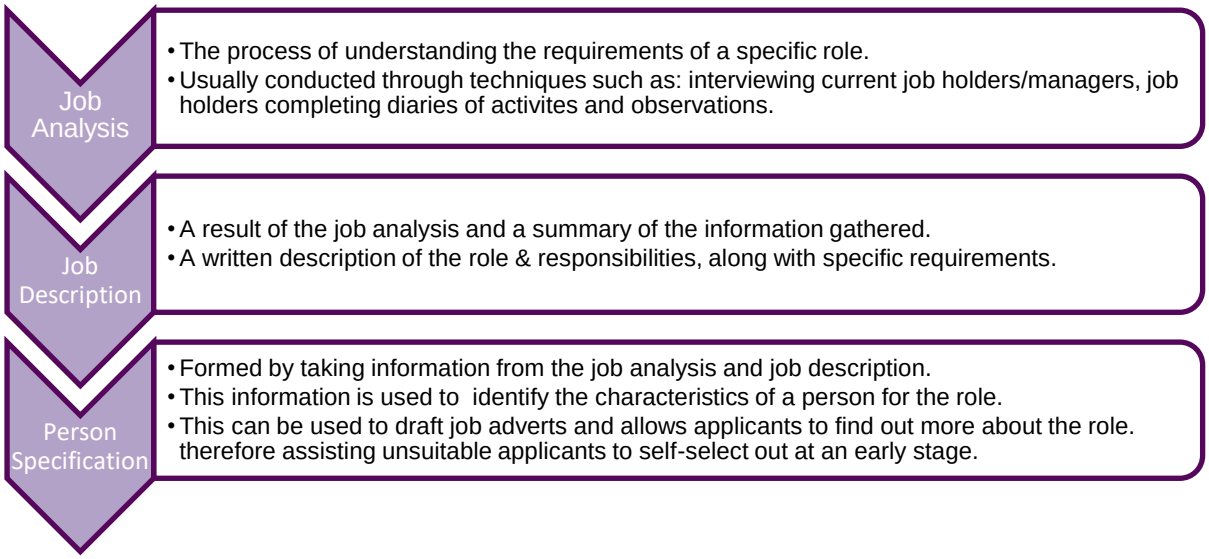
Objectivity – Psychometric tests provide an objective assessment of a person's abilities or other characteristics. If we try to make judgements about people it can be clouded by knowledge, biases and stereotypes. This objectivity makes the assessment fair to all people and all groups of people.

Accuracy – Also known as reliability, accuracy is a key feature of psychometric tests with considerable effort being made during the development process to ensure accuracy of testing. All measures contain a degree of error but in psychometric testing the existence of error is recognised and the degree of error is quantified. This means that test results are considered to be only an indication of a characteristic of a person (e.g. ability or personality) rather than being exact.

Validity – To be valid a test needs to be suitable for a specific purpose. Psychometric tests are developed to assess specific psychological constructs. Good tests provide evidence to demonstrate that they measure the construct they claim to measure. The research and development behind tests makes it easier to ensure their validity compared with some other types of assessment.

When should tests be used?

Before using any psychometric test, you need to have a clear understanding of **why** you are using it and **how** the results will be used. When using tests in an organisational context you must have a clear understanding of the role and therefore the requirements of the candidates. An established process for understanding role requirements and using this to identify selection methods is illustrated below:



Following this process, the tests must be selected based on the reliability and validity of the test, availability of relevant norm groups and the mode of delivery.

Intelligence and ability

A common question which surrounds ability tests is whether they are testing intelligence. There are problems using the term 'intelligence':

- Firstly, it is very emotionally loaded. To use a test to label someone as more or less intelligent may have strong connotations for some people.
- Secondly, it is hard to define what is meant by intelligence.

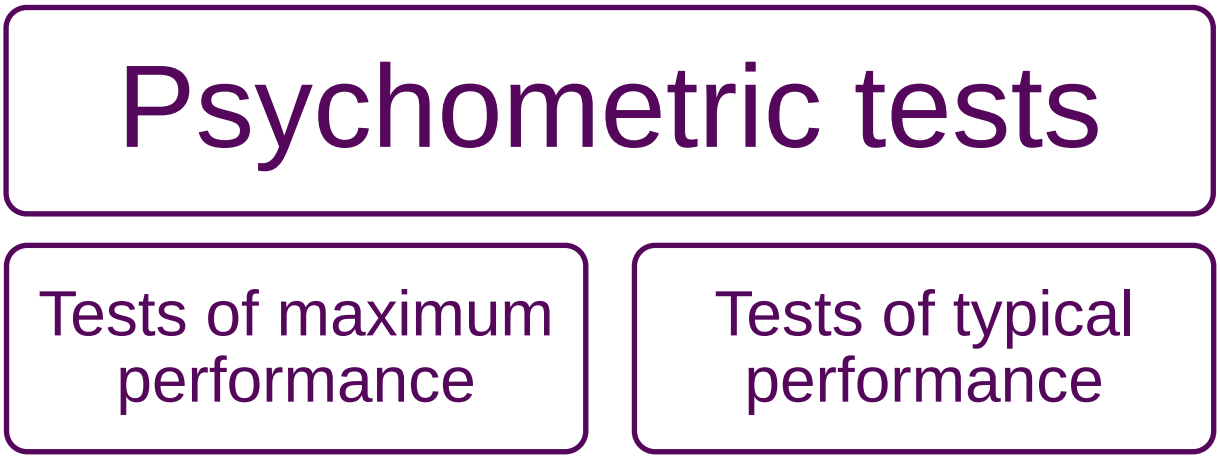
Over the past couple of decades many psychologists have come up with models or theories of intelligence. Gardner and Sternberg developed models which expand on traditional, more academic forms of intelligence, recognising that people can be intelligent in a number of different ways. For example, Gardner's (1983) seven intelligences include; linguistic, mathematical, spatial, musical, bodily kinaesthetic, interpersonal and intrapersonal intelligences, the latter two of which are linked to the recently popular concept which has been termed emotional intelligence.

This ambiguity surrounding what is meant by intelligence highlights the need for test users to understand what each test measures and which specific ability they relate to. Tests can be classified to help users identify what they measure.

By putting several of these specific ability tests together in a 'test battery' it may provide a broader understanding of a person's overall ability (their general ability), but will not give us an overall understanding of their intelligence.

Classifying tests

Thousands of tests have been developed and are regularly used by organisations. These tests assess a wide variety of psychological constructs, therefore test classification has been used to group similar tests to provide users with information about what they measure and how. We will go through the stages of classification; to begin with we will look at the first classification which is the distinction between tests of maximum and typical performance.



Tests of maximum and typical performance

There are three key characteristics which distinguish between these two types of test; these are detailed in the table below:

Characteristic	Maximum performance	Typical performance
Right / Wrong answers	✔	✘
The instructions to test-takers	'perform to the best of your ability in the time allowed'	'indicate how you usually or most typically behave, feel, react, etc.'
Time limit on the test	✔	✘

Example:

A maximum performance test question may look like this –

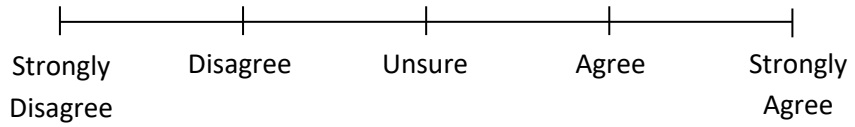
Which number is missing from this sequence 1, 4, 9, …, 25

a) 14, *b) 15,* *c) 16* *d) 17*

A typical performance test question may look like this –

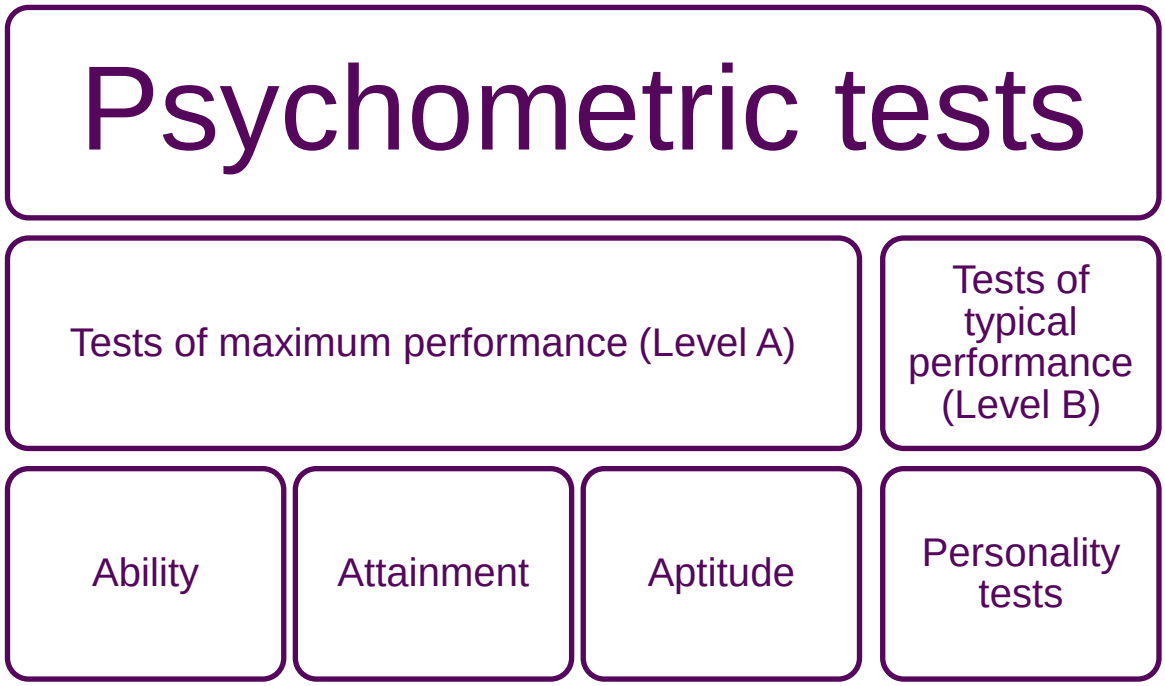
Please indicate the extent to which you agree with the following statement:

I get on with everyone that I meet.



The distinction between maximum and typical performance is often seen as reflecting the distinction between 'ability' and 'personality' tests, although there are other types of tests within these categories. This distinction is also captured in the BPS Test User: Occupational Qualifications, ability (Level A) and personality (Level B).

As you can see in the diagram below there is another layer detailing the distinctions within tests of maximum performance.

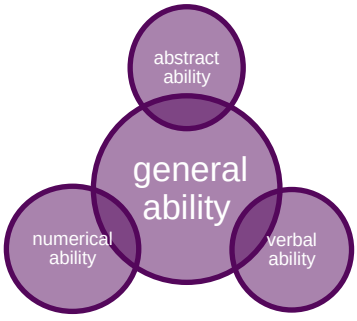


Ability tests

Ability tests are the most common form of occupational testing and are designed to measure an individual's cognitive ability by assessing what they are currently capable of doing, not just what they have learnt. These tests typically assess a type of reasoning through the use of verbal, numerical or abstract material.

Abilities can vary in how specific they are; therefore a distinction can be made between tests of general or more specific abilities. Work by prominent researchers such as Spearman (1927), Vernon (1950) and Carroll (1993) supports the distinction between general and specific abilities.

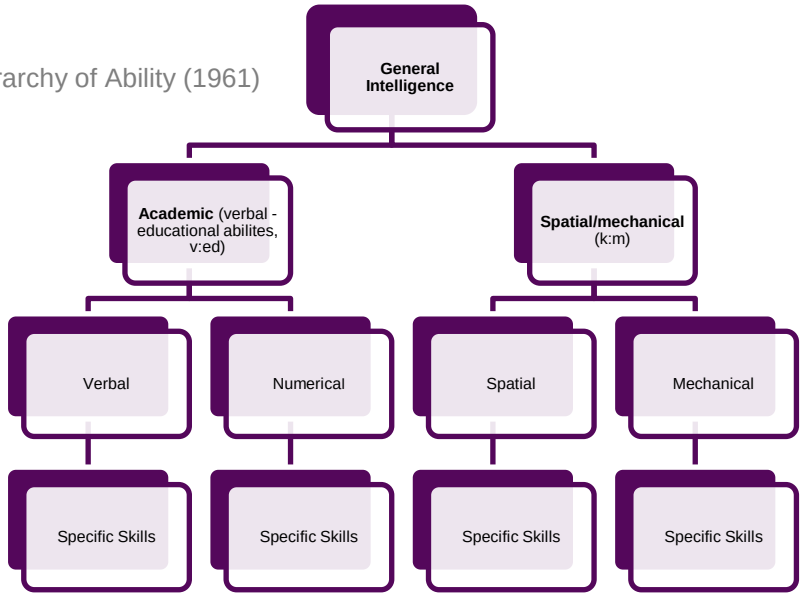
Spearman noted that two factors can account for a person's performance on any test; general ability and specific ability. The principle of 'general ability' came from the observation that if a person does well on one type of ability test, they are also likely to do well on other tests. As illustrated below, general ability contributes to performance on specific ability tests.



Whilst general ability will have an impact on performance on specific tests, like verbal, numerical and abstract tests, performance on each specific test will also be sufficiently independent of general ability for it to be meaningful to look at a person's pattern of inter-test scores.

Vernon (1961) and Carroll (1993) describe abilities in a hierarchy, with the common element of general ability at the top, followed by clusters of abilities in the middle layer and very specific abilities at the lowest level. Vernon's model is shown below; the second lowest level of this diagram is where many of the tests we will discuss during this course would be placed. The lowest levels relate to the specific skills which are assessed within these tests.

Vernon's Hierarchy of Ability (1961)



Through examination of Vernon's hierarchy, it is interesting to note that the academic grouping covers the types of abilities which are the basis of modern education-based tests and are abilities which are valued by western educational systems. However, the more practical abilities which are linked to the spatial-mechanical grouping are the ones which are traditionally seen as separate from the more 'academic' route.

Cattell (1971) proposed a theory of 'fluid' and 'crystallised' abilities, similar to Vernon's model as it too proposed a hierarchy with general ability at the top, but with fluid and crystallised abilities below.

> **Fluid ability –** to think in a flexible and creative manner, to learn new ideas and deal with novel situations. This is most commonly assessed using abstract reasoning tools.
>
> **Crystallised ability –** dependent on experience and learning and refers to the skills or knowledge acquired by an individual. With the exception of abstract reasoning tests, most ability and attainment tests fall under this category.

General and specific abilities

Abstract reasoning tests are considered to be measures of **general ability**, as they measure the ability to acquire new skills and learning across a wide range of areas, as well as dealing with novel situations. They use shapes as the stimulus material rather than verbal and numerical material so that they do not suffer from the effects of learning and educational experiences.

Verbal, numerical, mechanical and spatial ability tests on the other hand all measure more specific abilities. Often test users combine a series of these tests to form a 'battery of tests' to measure different specific abilities.

Attainment tests

Attainment tests are used to assess learned knowledge or acquired skills, e.g. educational qualifications, typing tests, tests of specialist knowledge (e.g. accountancy exams) and the driving test.

There is a strong association between attainment and ability as an individual's capacity to learn will be influenced by their ability. Attainment test results describe what knowledge or skills have been acquired through exposure to opportunities for learning. Unlike attainment tests, ability tests are not based on a defined body of knowledge or a syllabus although they do assume prior exposure to learning - for example, in order to complete a verbal test, the user must be able to read.

Aptitude tests

Aptitude tests are used to measure the mental abilities that affect the likelihood to acquire skills for a specific job. For example, a clerical aptitude test would not require knowledge of clerical procedures but would measure the mental processes required in order to acquire clerical skills. It would use tasks such as accuracy, filing, checking, etc.

The content of an aptitude test is often very similar to that of an ability test. It is possible to use an ability test for aptitude as the term aptitude is given to any test used to identify a person's potential to acquire skills. In this case, it is assumed that general ability with numerical and verbal material will predict performance in a job that contains these elements.

Alternative forms of occupational testing

Work sample tests

Work sample tests are an alternative form of occupational testing. Work sample tests involve specific aspects of a job being completed by applicants under standardised conditions. For example, a test for a clerical position may use a typing proficiency test for a specific computer software package. Work sample tests assume some level of proficiency in the area being assessed, making them an example of an attainment test. If applicants may not have the necessary skills or where they need to undergo training, an aptitude test is more appropriate.

When using a work sample test, the procedure and conditions should be standardised in order to compare the test scores. Once a task has been identified as relevant to the role, it can be replicated. During the development of a work sample test, it is important that the scoring procedure is clear with the aim to remove as much of the subjectivity as possible. By keeping the scoring objective, it helps standardise the scoring of the results.

These types of test provide a clear indication of an individual's performance on a given aspect of the role for which they have applied.

Personality tests

Overall job performance is a product of many factors including ability and personality. There are many different types of personality test which organisations can use to assess both potential and existing employees. Personality measures are used to assess someone's typical behaviour and preferences. For example, many personality tests measure 'conscientiousness'. A person who has a preference for working in a structured and organised environment would endorse these scales more and therefore have a high score on 'conscientiousness'.

Organisations can use personality information for selection and development. A detailed job analysis can be used to identify which areas of personality are relevant to the role. By comparing the personality profile to the job requirements it is possible to identify any strengths, potential concerns or areas for development.

Tests of motives and interest

Another form of psychometric testing which can be used by organisations investigates the motives and interests of their employees. Knowing what motivates an employee can be valuable to managers as they can use this to help make sure that all employees are motivated to perform to the best of their ability. They could use this information to tailor the roles to each individual based on their needs, interests and motivators helping to create a happy and more productive workforce.

The influence of environmental factors on ability

There is a large, politically and socially sensitive debate regarding the extent to which ability is innate or influenced by environmental factors. This debate is known as the 'nature-nurture debate' and results in research findings being open to interpretation. Most psychologists accept that both factors are important in the development of ability.

Research suggests that heritability accounts for over 60% of general cognitive ability (McClearn et al., 1997). This data is based on twin studies, which often use twins who have been reared apart. These studies found that these twins have surprisingly similar ability levels. Due to the common maternal environment before birth even the results of twin studies which uses those separated at an early age are not as clear-cut as is sometimes interpreted. With these figures it is important to note that there is still potential for significant environmental influences and highlights that ability is not unchangeable.

Many environmental factors have been studied in order to understand their long-term effect on ability. It is difficult to identify precise effects of the family environment as there are many possible factors which could be involved. One factor which has been found to influence performance on ability tests is having a stimulating environment where learning is encouraged and supported by parents. Education also impacts test scores but this is linked to the fact that many tests require a degree of knowledge in order to complete them.

There is a complex interaction between education and ability: the amount of education has been found to predict ability but ability also has been found to predict how long someone is likely to remain in education. This interaction most probably reflects the tendency for people who are more academically able to find more reward in academic environments and are therefore more motivated to increase their learning.

Short-term factors such as health, level of tiredness, motivation and anxiety also all have been found to influence ability test performance.

Section 2 - Introduction to Reliability & Validity

Introduction

Reliability and validity are two of the key properties of any psychometric test. Reliability refers to the extent to which the tool is accurate at measuring what it claims to measure. However, validity relates to the extent to which a test measures what it claims to and is therefore suitable for a specific purpose. Both of these properties will be introduced in this section.

During this section we will explore what reliability is, the role of error in measurement, the types of reliability and how it is measured. We will also define different kinds of validity and explore the importance of the evidence provided by each type. We will see how validity helps us choose the most appropriate tests for a given situation and how we can use validity evidence as a way of relating test scores to outcome measures. We will also explore the relationship between reliability and validity.

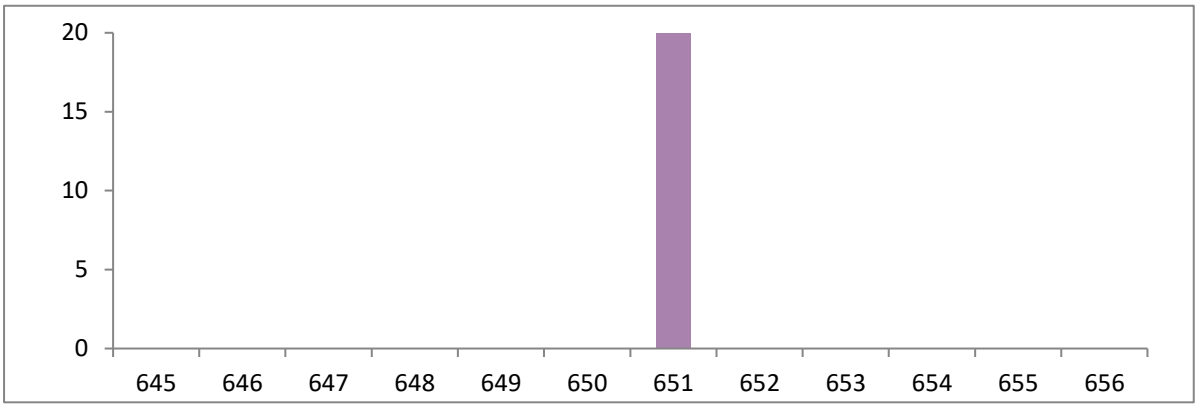
Reliability

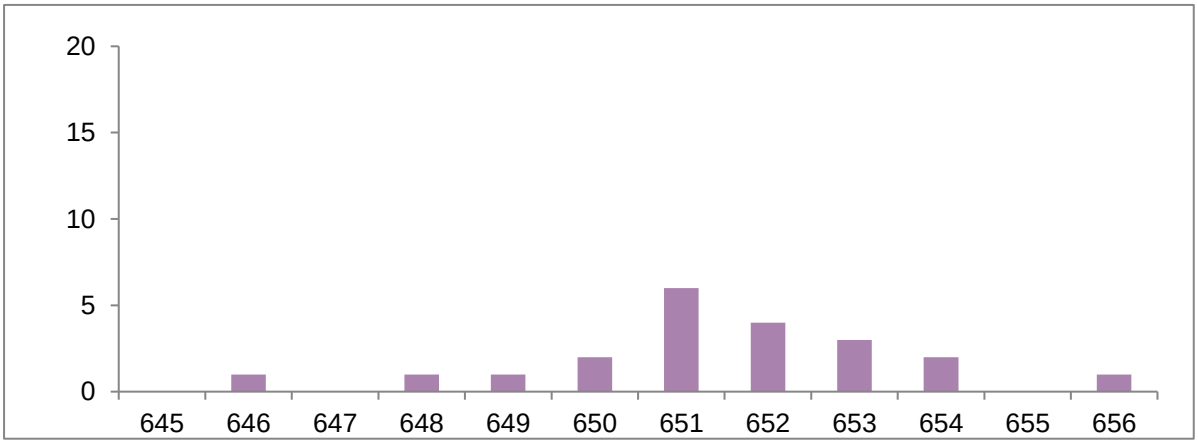
The concept of reliability refers to the accuracy of the measurements that we make. It is an essential property of any psychometric test and indicates the degree to which a person's actual test score is a result of their 'true score' on the construct being measured, and how much it is due to 'error'.

A practical example of reliability

Twenty people were asked to measure the length of the same table. They each use the same tape measure and are asked to record the length of the table to the nearest millimetre.
Would you expect all of the measurements they make to be exactly the same?

It is likely no matter how carefully the measurements are made that there will be some variation in the measurements taken by the 20 people. If all 20 measurements were completely accurate and so all the same, a histogram of the results would look like the one below. An actual histogram of the measurements is more likely, however, to look something like the second one.





What the second histogram illustrates is that there is general agreement about the length of the table with most people thinking its length is between 649 and 653. Only a few people measured it as being either longer or shorter than this. Therefore, we have a reasonable idea of the approximate length of

the table, even if we cannot be totally sure to the nearest millimetre. This highlights the issue of accuracy.

Accuracy

One of the main points in the theory of reliability relates to the degree of error which is contained within all measurements. This means that measurements cannot be assumed to be perfectly accurate. When dealing with physical objects it is easy to make small errors in measurement, however, the scope for error is greatly increased when measuring intangible psychological constructs such as ability, personality or motivation.

One of the dangers when using psychometric tests is to treat numerical results that are obtained as absolutes. Reliability allows us to quantify the accuracy of the test so we can treat the results in accordance to their accuracy. As qualified test users, we must not over-interpret test scores or make poor decisions based purely on the test results. We must recognise the degree of error inherent in any psychological measurement.

The result therefore can be expressed as the equation below:

Observed Score = True Score + Error Score

Error

Within the results of any measurement you have the true score and the error. Any test and testing session should aim to maximise the measurement of true scores and minimise the amount of error. This error can come from a number of sources, including:

Error due to the testing situation

Error from the individuals taking the test

Error in the test itself

Error due to the testing situation

When administering any test, it is important to ensure that the test conditions are as similar as possible for all test-takers. This standardisation is one of the essential properties of psychometric tests and should mean that all test-takers receive the same instructions, have the same time to complete the test (assuming it is a timed test) and that all test-takers complete the test under appropriate conditions. Variations between test users (e.g. deviation from standardised instructions) and testing environments (e.g. taking the test in a room that is too hot or subject to excessive outside noise) can be a source of measurement error, as will any inaccuracies when scoring tests.

Error from the individuals taking the test

An individual's performance can be influenced by a number of possible factors. For example, test-takers may vary in their mood, motivation, level of anxiety and physical well-being. Giving guidance to

test-takers prior to attending the test session, coupled with good administration, can help minimise error resulting from individual test-takers. It also helps gain 'buy-in' from the candidate.

A test-taker's prior experience can affect their performance. On average, those who have completed psychometric tests previously are likely to show an increase in their score compared to those test-takers who have not. These improvements are as a result of increased comfort and familiarity that comes with experience, and the opportunity to use prior experience to develop test-taking strategies. There could also be a memory effect that influences their score. One of the main purposes of test preparation materials is to compensate for these differences in prior experience.

Error in the test itself

No test can ever be a perfect measure of the psychological construct or constructs it claims to measure. Factors such as the length of the test, the adequacy with which the test covers the domain it assesses, the chances of guessing a question correctly and any ambiguities in the test questions all contribute to error in the test. Reputable test publishers follow systematic development processes to ensure that test error is kept to an acceptable level, though error will never be eliminated altogether.

The formula which is expressed above highlighting the role of error in test results is part of Classical Test Theory. This theory suggests that whenever a score is obtained from a test the 'observed test score' should be viewed as having the true score and an error component. Classical Test Theory recognises that error is inherent in all measurements we make when using tests.

Test theory also tells us that error is randomly distributed. Therefore, the error component in any test means that an observed score may be an underestimate or overestimate of a person's true score. As with many other aspects of psychometric testing, this distribution of error is assumed to be normally distributed around a person's observed score.

Theories of test measurement

Classical Test Theory

Item Response Theory

Classical Test Theory

Classical Test Theory has been the main theory underpinning test measurement over the past eighty years. It assumes that the observed score produced by a test is made up of the participant's true score and some degree of error. With Classical Test Theory, ability is estimated by counting the number of items answered correctly.

The score is test-dependent and it cannot be assumed that the differences in scores are equal; you also cannot conclude that the difference between a score of 12 and 13 is necessarily the same difference as between a score of 21 and 22.

Item Response Theory

Item Response Theory (IRT) is an alternative approach used to gather information about an individual's ability.

IRT looks at each individual question (or item) and its level of difficulty. A test based on IRT adjusts the questions given to the test taker based on their performance. It uses information about the difficulty of each question in order to accurately differentiate between ability levels.

Thanks to IRT, test publishers have been able to develop and maintain item banks, large banks of items relating to a given ability, and create computerised adaptive testing. The flexibility allowed by Item Response Theory over Classical Test Theory means that tests can be adapted to include harder questions based on the participant's performance. As tests based on Classical Test Theory use a standard set of questions for all candidates, this kind of adaption to the questioning is not possible. Publishers of tests using Classical Test Theory need to develop tests at different levels of difficulty to distinguish between candidates with different abilities. However, with Item Response Theory, a single test can differentiate between people of broader ability levels.

Tests developed using Classical Test Theory calculate the candidate's ability by assigning one point to each question answered correctly, irrespective of each question's difficulty level. However, IRT enables us to measure a candidate's true ability more accurately and quickly because each item is weighted by its difficulty level. In this way, tests based on IRT are more reliable.

Example

A numerical reasoning test has ten questions at different levels of difficulty:

Item number	Item	Difficulty Level
Question 1	2 + 2	Easy
Question 2	8 – 3	Easy
Question 3	5 + 1	Easy
Question 4	3 x 3	Medium
Question 5	4 x 12	Medium
Question 6	36 ÷ 6	Medium
Question 7	$\sqrt{(3^3 \div 5)}$	Hard
Question 8	(45 x π) ÷ 3	Hard
Question 9	$\frac{(3\times9)^2}{12^3}$	Hard
Question 10	$S_k = (-2)^k \int_0^1 log^k\ (sin\pi x)dx$	Really Tricky

When the candidates take the test, the computer will adapt the questions they answer based on their ability - this technique is known as **adaptive testing**.

Let's imagine there are two candidates taking the test.

Candidate one - the first item is question 1 which the candidate answers correctly. The computer will next present a medium-level item, question 4. If the candidate answers this correctly, the computer will next present a hard-level item, question 7. However, if the candidate answers question 4 incorrectly, then the computer will present another medium-level item, e.g. question 5.

Question 1 ✔

Question 4 ✔

Question 7 ✔

Candidate two - the first item is question 1 which the candidate answers incorrectly. The computer will next present another easy-level item, question 2. If the candidate answers question 2 correctly, the computer will next present a medium-level item e.g. question 4.

Question 1 ✘

Question 2 ✔

Question 4 ✔

With the computer continually adjusting the questions selected based on each candidate's test performance, we gain a measure of the candidate's ability more quickly and accurately compared to tests based on Classical Test Theory. By being able to adapt the questions depending on the ability level of the candidate, we obtain more information about each candidate's actual ability. The amount of information provided by a test about a candidate's ability is termed the **test information function**.

Methods for measuring reliability

Reliability of a test is determined in practice by looking at its consistency. There are three methods used to measure reliability.

Internal consistency

Internal consistency relates to the consistency of the test items as a whole by estimating the extent to which the different parts of the test are measuring the same construct. To evaluate this, the test is broken into separate parts and the correlation between these parts is calculated.

One way of splitting the test is dividing it into two and calculating a score for each part. The two scores from a number of test-takers are then correlated to determine the strength of the association between them. This method is known as 'split half', and most often the test is split according to odd and even numbered test items.

One method more commonly used by modern test developers is known as Cronbach's Alpha. This approach takes each test item in turn and examines the association between the individual item and the total test score minus that item. A single statistic is then produced to summarise the extent to which each individual item is associated with the total score on average.

Test-retest reliability

Test-retest reliability relates to the consistency of test scores over time comparing test results from time one and those at time two. Calculating test-retest reliability requires that one group of test-takers complete the test twice, with a sufficient gap between them to allow for some fall-off in test-takers' memory for the test items. There are no set rules for how far apart these two test times should be but it is unlikely to be less than two to three weeks or more than twelve months.

Parallel form

Parallel form is the consistency of the scores produced from two versions of the same test; these tests would have been developed by the test publisher in parallel to assess the same construct. A group of test-takers would complete both tests and the results from both tests would be correlated. A good parallel form reliability study would involve half the test-takers taking the first version of the test then the second, whilst the other half take the second version and then the first.

Comparing different methods for estimating reliability

Each of the three types of reliability involves a different procedure for gathering the data on which they are based. Internal consistency is the simplest to collect data for as it requires only a single administration of a single test. Because of this, virtually all test manuals will include an estimate of internal consistency reliability.

It is more difficult to collect data for both test-retest and parallel form reliability studies because test-retest reliability relies on getting the same people to sit the test twice and parallel form reliability requires the publisher to develop two versions of the same test. Motivation of the test-takers may be a factor, especially for test-retest reliability studies when they are taking exactly the same test for a second time with little reason other than to provide data for the test publisher.

Though test publishers may not produce two versions of the same test necessary for parallel form reliability, the move to computer-delivered tests means that different versions of a test may be available in some cases. Some modern computer-delivered tests are based on ‘item-banking’. Item banking involves generating a large bank of items from which different versions of a test can be automatically compiled. Such tests are designed to reduce problems with item exposure and familiarity, and go some way to ensuring the authenticity of unsupervised tests. If item-banked tests are available, parallel form reliability can be established using such a system.

Validity

Whenever we look to use a test or method of assessment, validity should be our ultimate consideration. We will come to see that when selecting a test, it is important to ensure it meets all of the psychometric criteria outlined in Section 1. These properties, such as reliability and having a suitable norm group, are important for test selection but they also help contribute to the validity of a test.

Essentially, validity answers the question 'Is the test suitable for my purpose?' Assuming that the necessary psychometric properties are in place, we need to seek evidence for the validity of the tests or other assessment methods we are considering using. Validity evidence can come from a number of sources, with theses commonly being known as: face validity, faith validity, content validity, construct validity, criterion validity and consequential validity. In judging a test's validity, it is common for users to look to a number of these areas for evidence before making an overall judgement about the test's suitability.

In one key aspect, validity is very different from the other properties of tests. Many of the technical properties of a test should be established by the test developers before the test is published, including some of the initial evidence for the test's validity. This means, for example, that providing the test is administered under standardised conditions we can assume that the reliability values quoted in the test manual apply to our use of the test. As validity is about the suitability of a test for a specific purpose, validity is not a fundamental property of the test in the same way that reliability is. **Validity needs to be established each time a test is used for a new purpose.**

Identifying the criteria for validation

Understanding the job role and its requirements leads to a consideration of which tests or assessment methods should be most appropriate for assessing the different aspects of job performance. These requirements need to be revisited when conducting a validation study as they form the basis of the criteria.

It can be difficult to define good performance but this is what tests should be validated against. A common problem for criteria validation is that job descriptions developed by organisations don't always detail the performance measures for a role. They commonly refer to what a person is required or expected to do but don't necessarily define good performance.

Concurrent and predictive validity

Criterion validity involves analysing the association between test scores and some aspect of job or training performance. This can be done using the statistical method of correlation; this will be explained later on during the course.

Concurrent and predictive validity are methods of testing the criterion validity and will be explained during the measuring validity section of the course.

The six types of validity

As stated above there are a number of sources which can tell us about the validity of a test. There are in fact six sources of validity, these are illustrated and explained below.

Face validity

Face validity is the extent to which the test-taker perceives the test to be measuring something relevant to the purpose of assessment. It is not quantifiable, but is important as it affects the test-taker's motivation to complete the test. Good administration can enhance the face validity by explaining what is being measured and how it relates to the purpose of assessment.

Faith validity

Faith validity refers to the administrator's belief in the test and how much they believe the test is measuring something relevant to the purpose of the assessment. This type of validity doesn't require any evidence to support the notion that the test is valid. Although this is independent of the test takers beliefs in the test, a test with high faith validity is also likely to have high face validity.

Criterion validity

Criterion validity is the link between test scores and external criteria such as performance measures. The criteria we are interested in predicting may be specific for a particular application so criterion-related validity should be established for each new use of a test. By investigating the relationship between performance measures and test scores you can begin to predict likely job performance or ability of applicants to acquire the skills necessary for the job. Due to the link between external criteria and test scores it is thought of as the most important type of validity.

Construct validity

A major source for the evidence of validity as it is based on empirical research. Statistical evidence determines the extent the test measures what it claims to. This can be done by observing the correlation between the scores on this test and scores on other tests, or by looking at differences between groups of people who would be expected to differ on the construct assessed by the test. If the correlation between scores on two tests is high, it can be assumed that they measure similar constructs. Construct validation is an on-going process.

Content validity

Content validity requires a close relationship between the content of the test and the content of the job.

A clear definition of the construct to be measured can lead to a framework used in item development. Item trialling ensures that the items are functioning statistically and there is content validity.

When it is difficult to define the construct, subject matter experts can be used to rate the relevance of items.

The link between the content of the job and the content of the test should be based on a job analysis.

Consequential Validity

Consequential validity refers to the appraisal of value implications and the social impact of the score's interpretation. This includes both positive and negative consequences of the scores and both intended and unintended consequences of the scores. For example, there may be particular bias towards certain groups or individuals with certain characteristics.

Reliability & Validity

Reliability and validity are key properties of a psychometric test and are linked. Test publishers test both reliability and validity and these are published in the manual. If the test is administered in a standardised way, the results will be as reliable as the manual claims. Standardised administration helps remove some of the error which can be added through administration, e.g. variance in the detail of the instructions given to candidates.

It is important to note that a test cannot be valid without being reliable, although it is possible to be reliable and not valid. The test must have both properties to be classified as a psychometric tool.

The reliability confirms the accuracy of the tool and the validity confirms that it is measuring the constructs which it claims to measure.

For a test to be accepted as reliable, it must meet certain statistical criteria. These criteria relate to the value of the reliability coefficient and will be covered in more detail later on in the course.

Section 3 – Best Practice

Introduction

When you conduct testing, it is important to follow the principles of best practice. It is a legal requirement to treat people fairly. Within this section we will cover the importance of having a testing policy and to undertake fair testing.

Fair testing

The Equalities Act (2010) states that you must make reasonable adjustments for candidates, either to the tests themselves or to the testing venue. The Act also provides information on what information should be protected.

When you test an individual, you have a responsibility to ensure that all personal information, including test scores, are kept securely with limited access. The test-taker must be informed about what will be done with their results and who will have access to them. As the test-taker, they are the owner of their results and have the right to request access to them. There may also be a shelf-life for the test scores. Best practice would usually recommend destroying test scores after a period of 12-18 months. This must be done in a secure way as the test scores are classified as personal information.

To ensure that all members of the organisation adhere to the principles of best practice, it is useful to have a psychometric testing policy. Below we have detailed the key areas to cover within a psychometric testing policy along with some examples of how it would be written.

By making the guidelines clear it is easier for all employees to conduct testing in the same way following the principles of best practice. It also means that all test-takers will be treated equally.

The British Psychological Society is the governing body which oversees the use of psychometric testing in occupational settings and requires test users to follow the ethical guidelines they have set out. For more details, see: *http://www.bps.org.uk/what-we-do/ethics-standards/ethics-standards*

The Equalities Act

The Equalities Act (2010) ensures the protection of certain characteristics of individuals in order to prevent any form of discrimination. The protected characteristics include: age / disability / gender reassignment / marriage and civil partnership / pregnancy and maternity / race / religion or belief / sex / sexual orientation.

Two of the forms of discrimination include direct and indirect discrimination.

Direct discrimination relates to when an individual is being treated less favourably than others due to one of the protected characteristics. The law includes the comparator to the treatment of others as the law is related to equality and specifically fair treatment.

Indirect discrimination relates to the application of a provision / criteria / practice which is a discriminator in relation to one of the protected characteristics of an individual. The law covers:

- Those who experience the application of a provision / criteria / practice when they do not share that characteristic.
- Those who share a characteristic and have been put at a disadvantage when compared to others who do not share it.
- Those who are put at a disadvantage as a consequence of applying a provision / criteria / practice.
- Those affected when the application cannot show that it is a proportionate means of achieving a legitimate aim.

The Equality Act (2010) aims to protect disabled people and prevent disability discrimination. The act defines a disability in this way: 'a person has a disability if they have a physical or mental impairment, and the impairment has a substantial and long term adverse effect on their ability to perform normal day–to-day activities.'

An employer must not directly discriminate and must not indirectly discriminate without a fair and balanced reason.

Reasonable adjustments

An employer is obligated to make reasonable adjustments; these should avoid disabled candidates being put at a disadvantage compared to other candidates.

These adjustments may include any working arrangements or physical aspects of the workplace such as:

- Physical adjustments to the workplace
- Allocating some duties to another employee
- Moving a disabled employee to a different but suitable job
- Altering hours of work
- Moving a disabled employee to another place of work
- Allowing time off during working hours for treatment or rehabilitation
- Arranging training for the employee
- Acquiring or modifying equipment
- Altering instructions or reference materials
- Altering procedures for testing or assessment
- Providing a reader or an interpreter
- Providing supervision

As an employer you cannot issue pre-employment questionnaires before offering employment. Health related questions may only be asked prior to offering an individual a job to:

- Help decide whether there is a need to make reasonable adjustments to the selection process
- Help decide whether an applicant can carry out an essential part of the job
- Monitor diversity amongst applicants
- Take positive action to help disabled people

The effect of disabilities

Having a disability may affect the results as adjustments may need to be made for the test-taker which could impact the reliability and validity of the test. Any changes to the administration outlined by the test publisher will affect the overall reliability of the results. It is important to make adjustments following consultation with the test publisher in order to try and keep the administration as standardised as possible in order to reduce the effect on the reliability of the results.

Adverse Impact

When testing, there is the potential for an adverse impact when there is a difference in the performance of two groups such as gender. If there are significant differences in average performance between groups, this may impact the selection rates. If a subgroup is selected at a rate of less than 80% of the group with the highest selection rate, this would indicate adverse impact. No form of people assessment is immune from unfair bias. The UK Equal Opportunities Commission (EOC 1988) state that research findings have been able to show that aptitude tests are more valid and fair than many other types of assessment for selection for both the US and UK.

If the group with the lower selection rate are covered by employment law, then adverse impact can be grounds for litigation. However, the differences between groups may relate to real differences in job performance levels therefore the discrimination is fair.

Unlike with indirect discrimination, adverse impact is when it is justified to use processes or procedures but they result in a difference in performance. For example, if you use a test which is valid for your purpose but the selection between groups has a difference it is most probably due to a true difference in performance and not discrimination.

Psychometric Testing Policy – Themes & Examples

The policy contains the following sections:

- A mission statement
- Responsibilities for testing standards
- When tests are used
- How tests are chosen
- Commitment to equal opportunities
- How test scores will be used
- Confidentiality & storage of results
- Responsibilities to test takers
- Monitoring of testing
- Copyright implications

1 Mission Statement

Ensures readers are aware of the objectives of using tests and why a policy exists. Thus it may cover the applications for which tests will be used, why tests are being used and the importance of upholding the standards laid down by relevant professional bodies with regards to testing.

Example:

We use psychometric tests to enhance the quality and quantity of information available for selection, development and training decisions and as an aid to organisational change. We are committed to the highest standards of practice in the use of all psychometric tests in order to maximise the benefit to the organisation and the individual and to promote fairness and equality of opportunity for all.

2 Responsibilities for Testing Standards

Responsibilities and accountabilities of each individual to be involved in psychometric testing must be clearly defined. Instruments are potentially dangerous in the wrong hands. Thus the policy statement should include clear guidelines on the specific training required at each level and each stage of the testing process - i.e. admin / interpretation / feedback of different types of instrument.

Example:

Each test user must ensure that he / she uses tests to the highest professional standards and only in accordance with the guidelines set out in this policy.

Only trained test users who hold the relevant qualifications may use and interpret psychometric instruments. Trained test users may delegate test administration to a person trained in this area.

3 When Tests Are Used

The policy must show evidence of evaluating different situations to establish whether it is appropriate to use tests for different occasions - i.e. will a test provide additional information and add value in selection / development / counselling / team building. Important to ascertain at what part of the process should tests be used, in conjunction with what other methods and how the results of the tests fit into the overall procedure (i.e. how are they used).

Example:

Tests may be used for selection, development and counselling purposes. Any additional uses should be referred to the central testing unit for approval.

4 How Tests are Chosen

Make it clear how tests are to be selected (job analysis to ensure tests measure skills relevant to the job; thorough consideration of the reliability, validity, acceptability, norms, availability of the test). It is also important for the statement to identify who is responsible for selecting appropriate tests.

Example:

All psychometric tests used must be clearly relevant to the given purpose. Detailed job descriptions and person specifications based on objective job analysis must be prepared prior to the choice of tests for any selection or promotion procedures. All decisions to use tests should be clearly documented and copies should be stored in an agreed central unit.

5 Commitment to Equal Opportunities

Tests give objective information about a candidate and have been shown to lead to better and fairer employment decisions. On occasion however, tests have sometimes been found to have a disparate impact on gender / ethnic groups. Awareness of this means that guidelines are followed to avoid inappropriate use of tests. These should be clear in a policy statement. Consideration should be given to these issues when selecting tests and in monitoring outcomes of procedures - i.e. select on merit, using measures which are clearly relevant to the job demands and which are free from extraneous bias. Similarly, candidates with disabilities must not be discriminated against.

Example:

The organisation is committed to selection on merit and only measures which are clearly relevant to job demands and are free of extraneous bias should be used. All assessments for selection and promotion must be monitored to ensure they do not unfairly exclude or disadvantage any section of the population.

6 How Test Scores will be used

Tests should always be interpreted by properly trained individuals in the context of clearly defined and relevant employment criteria. Raw data should never be given to people who are not fully conversant with the instrument or the meanings of particular scores.

Individuals must be trained so that they are aware of the limits of the information and do not over interpret it. It is also important that interpreters are trained so that they are aware of the importance of selecting appropriate, relevant comparison groups.

It is vital that the policy statement stresses the importance of pre-defining criteria which has been identified as essential / desirable for job performance and of sticking to these criteria throughout any decision making process.

Example:

Test scores must be interpreted by suitably qualified personnel on the basis of relevant norm groups. Fixed cut-offs may only be imposed where specific evidence of test relevance is available (e.g. job analysis, validation study).

7 Confidentiality & Storage of Results

Test materials should not be allowed to circulate freely, and must be kept under lock and key at all times. Respondents should never be allowed to take materials away with them nor should any testing be conducted by post. Test results must be stored with due regard to confidentiality. Access restricted to those with a need to know (those involved in the process) and with those whom the respondent has agreed to share the results with in admin or feedback. Psychometric data can become less accurate over time. Thus scores should not be kept on file indefinitely - time periods for which scores are valid depends on the instrument and individual circumstances (typically 6-12 months).

Example:

Test users must ensure that all test materials are securely stored. Test results should be kept by test users in locked files. A written interpretation of results should be kept in personnel files and provided to relevant individuals. Results over 12 months old are invalid for selection or promotion decisions. All results are to be destroyed after three years or when the respondent ceases to be employed, whichever is sooner.

8 Responsibility to Test Takers

It should be made explicit what pre-briefing candidates will get, the responsibility of the administrator at the testing session and of the person giving feedback to the candidates. It is important that testers are open and honest about why tests are used, how results will be used and the arrangements for feedback. What the tests measure and how they are relevant to the process must be explained in relation to job-related criteria. Selection feedback can be of significant benefit to unsuccessful candidates and can leave them with a positive view of a negative outcome. Face-to-face feedback is preferred but telephone feedback may be more practicable.

Example:

This organisation is committed to dealing fairly with all candidates to be tested. We will be open and honest about the use of tests, provide suitable practice materials and relevant feedback whenever tests are used.

9 Monitoring of Testing

The use of psychometric instruments should be continually monitored to ensure continued appropriateness and effectiveness. Need to ensure that techniques used remain relevant to the job and that the criteria that you are assessing against has not changed. Monitoring by ethnic group, gender and disability is required to identify any adverse impact.

10 Copyright Implications

The reproduction of test materials without the permission of the author is a criminal offence, whether or not the reproduced materials are to be sold. Illegal copying of materials leads to lack of standardisation and poor control of materials, and gives respondents a bad impression. Ultimately, the resulting loss of income will lead to less new test development or higher prices.

Example:

Under no circumstances should any test materials be photocopied or installed on computer without the test publisher's permission.

Section 4 – Describing Test Scores

Introduction

To make the test scores meaningful they need to be interpreted in a standardised way. We will start by looking at how to describe groups of test scores using measures of central tendency and spread, which is important when interpreting test scores in relation to a broader comparison group of scores. We will then introduce different types of distributions and their properties.

When a test is completed, a candidate's score is termed the 'raw score'. On their own, raw scores have very little meaning.

Maximum performance tests – these raw scores are usually the number of questions a person has answered correctly.
Typical performance tests – these raw scores will indicate the extent to which the respondent has endorsed items relating to a particular scale.

Example:

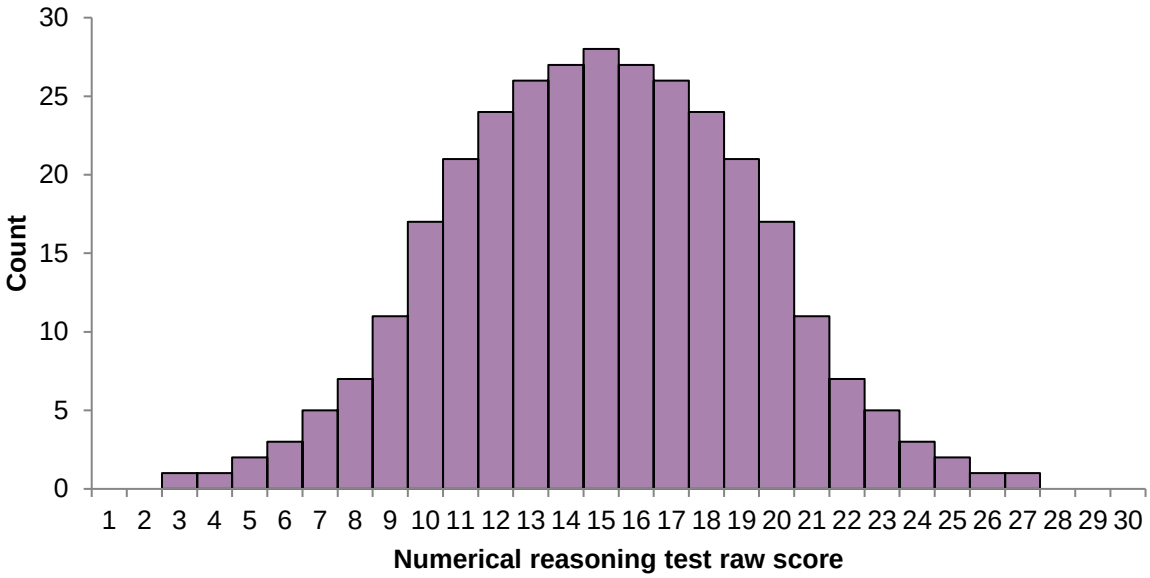
Jessie answered 17 out of 30 questions correctly on a numerical reasoning test. Would we interpret from this that Jessie has had a good level of numerical reasoning ability or maybe that it is about average? Depending on how difficult the numerical reasoning test is and how other people perform on the same test, we may even conclude that Jessie has not performed very well.

In order to understand a test score, it is necessary to know how people tend to perform on the test. If we know that most people who take the numerical reasoning test tend to score between 11 and 16, then we are likely to have a very different interpretation of Jessie's score than if we know that most people score above 18.

By understanding how people typically score on tests, we can make meaningful interpretations of each test-taker's score.

Score distributions

We tend to make sense of test scores by comparing them to people who have already taken the test. This is known as ‘norm referencing’ and the groups of people who we compare test-takers against are known as ‘norm groups’. A typical norm group may look like the one below.



This chart displays the scores of 306 people who completed the test. There was a possible raw score from 1 to 30; this is shown on the horizontal axis with the number of people obtaining each of the possible scores, or frequency, on the vertical axis. The technical name for this type of chart is a ‘frequency histogram’.

To help us interpret the test-taker’s score we can use a histogram. For example, we can see that a score of 8 would be quite poor as many people score higher than this, but a score of 22 would indicate a strong performance as only a few people achieve a higher score. Going back to our example of Jessie who scored 17 out of 30 on the numerical reasoning test, we can use the histogram of test scores to interpret her performance. A score of 17 places Jessie clearly above the mid-point, indicating that her performance is above average.

Histograms provide a convenient way of visually summarising the test data from a group of respondents. There are also numerical methods of summarising sets of test scores, which describe the distribution in terms of central tendency and spread.

Measures of central tendency

Measures of central tendency are ways of indicating the mid-point or middle value in a sample of test scores. The methods that can be used to do this are known as the Mean, Median and Mode.

The Mean: also indicated as $\overline{X}$, is the arithmetic average of a sample of scores and is the most widely used measure of central tendency. It is found by adding up all the scores to come to a total value, then dividing this value by the number of observations.

The formula for the mean is:

$$\bar{X} = \frac{\sum X}{N}$$

Where:

$\bar{X}$ = Mean
X = Raw score
$\sum$ = Sum of
N = Number in group

Example:

Person	Test score
1	11
2	17
3	18
4	9
5	25

Sum of scores	$\sum$ X	80
Number in group	N	5
Mean	$\overline{X}$	16.0

The Median: The middle value in a set of scores. To find the median, the data first has to be ordered from lowest to highest in value, and the middle value then found. Using the following data, ordered from the smallest to the largest value we can calculate the median:

Example:

10, 12, 12, 14, 15, 17, 18, 19, 25

In this case, the median is the middle number... 15.

In cases where there is an even number of numbers, the median will be half way between the 2 central numbers.

10, 12, 12, 14, 15, 17, 18, 19, 25, 28

In this case, the median is half way between 15 and 17... 16.

The Mode: The most commonly occurring value in a set of data.

Example:

10, 12, 12, 14, 15, 17, 18, 19, 25, 28

All values occur once except for 12 which occurs twice. As 12 is the most commonly occurring value, it is the modal value for this data set.

When a set of data has only one mode the distribution is described as ‘uni-modal’. If a set of data has two modes, it is described as ‘bimodal’.

Measures of spread

Measures of spread are ways of indicating how widely dispersed a set of data is. The methods most commonly used to describe the spread of data are the Range and Standard Deviation.

The Range: The distance between the highest and lowest values in a set of data is known as the range.

Example:

10, 12, 12, 14, 15, 17, 18, 19, 25, 28

The lowest value is 10 and the highest value is 28. Subtracting the lowest score from the highest (28-10) gives a value for the range = 18.

The Standard Deviation: The most widely used method of calculating the spread of a set of data. It is a comprehensive method which takes into account of how far every value in a set of data is from the mean. There are two versions of the formula for the standard deviation: the standard deviation for a population and the standard deviation for a sample, as shown here:

The formula for Standard Deviation of a population:

$$SD = \sqrt{\frac{\sum(X-\bar{X})^2}{N}}$$

The formula for Standard Deviation of a sample:

$$SD = \sqrt{\frac{\sum(X-\bar{X})^2}{N-1}}$$

The formulae are very similar, the only difference being the number at the bottom (the denominator) where this is *N* for the standard deviation of an entire population and *N-1* for the standard deviation of a sample. The -1 is the adjustment for the fact that samples tend to slightly underestimate the true standard deviation of the population. Most studies use a sample (e.g. a sample of managers) as it is very rare that you will have data from a whole population (e.g. every manager in the country).

Example calculation:

Person	Test Score	$X - \overline{X}$	$(X - \overline{X})^2$
1	17	-3	9
2	19	-1	1
3	20	0	0
4	20	0	0
5	21	+1	1
6	23	+3	9

The Mean $\overline{X}$ = 120 ÷ 6 = 20

SD = $\sqrt{\frac{\Sigma(X-\bar{X})^2}{N-1}}$

SD = $\sqrt{\frac{20}{5}}$ = $\sqrt{4}$ = 2

Using the mean and standard deviation

Knowing the mean and the standard deviation allows us to describe the 'shape' of a set of scores. The mean indicates where the mid-point of the set scores lies, but without the standard deviation we have no way of knowing whether all the scores are closely bunched together or spread out. Similarly, knowing t1he standard deviation tells us whether the scores are bunched together or spread out, but does not tell us where the mid-point of the distribution lies. So both the mean and the standard deviation are necessary to describe a set of test scores.

The normal distribution

Also known as ‘the bell curve’ due to its shape, the normal distribution is an example of a special type of distribution. It is a symmetrical distribution reflecting the tendency for observations to cluster around the mid-point and fall off towards the ends or ‘tails’ of the distribution.

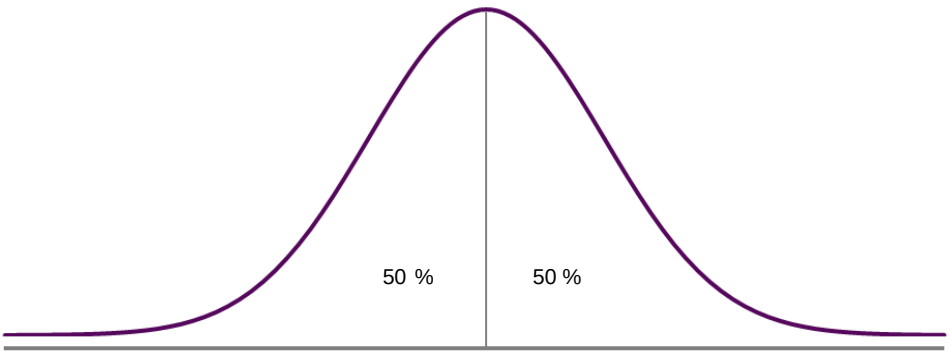
It is an important model as it is shows the distribution of many naturally occurring observations in the natural and social sciences. In the natural world, measurements of height, weight and length from the same species all tend to follow the normal distribution (e.g. the weight of all female humans, the heights of plants from the same species).

The normal distribution is important in psychometrics as the distribution of many psychological constructs match the normal distribution sufficiently closely for it to be used as a model for interpreting test scores. For example, aspects of ability and personality are assumed to be normally distributed in the general population: the majority of people will fall towards the middle on such constructs, with fewer people being very high or very low on them.

As the statistical properties of the normal distribution are known, and as many of the constructs we are interested in measuring through tests match the normal distribution, it can be very useful when interpreting test scores. Knowing the mean and standard deviation of any set of test scores that are approximately normally distributed, allows us to understand how many people are likely to have scored higher or lower than any given test score.

Properties of the normal distribution

In a normal distribution, the mean, mode and median all fall at the mid-point of the distribution and, as it is perfectly symmetrical, 50% of observations will fall to either side of the mean. These features of the normal distribution are illustrated in the diagram below.



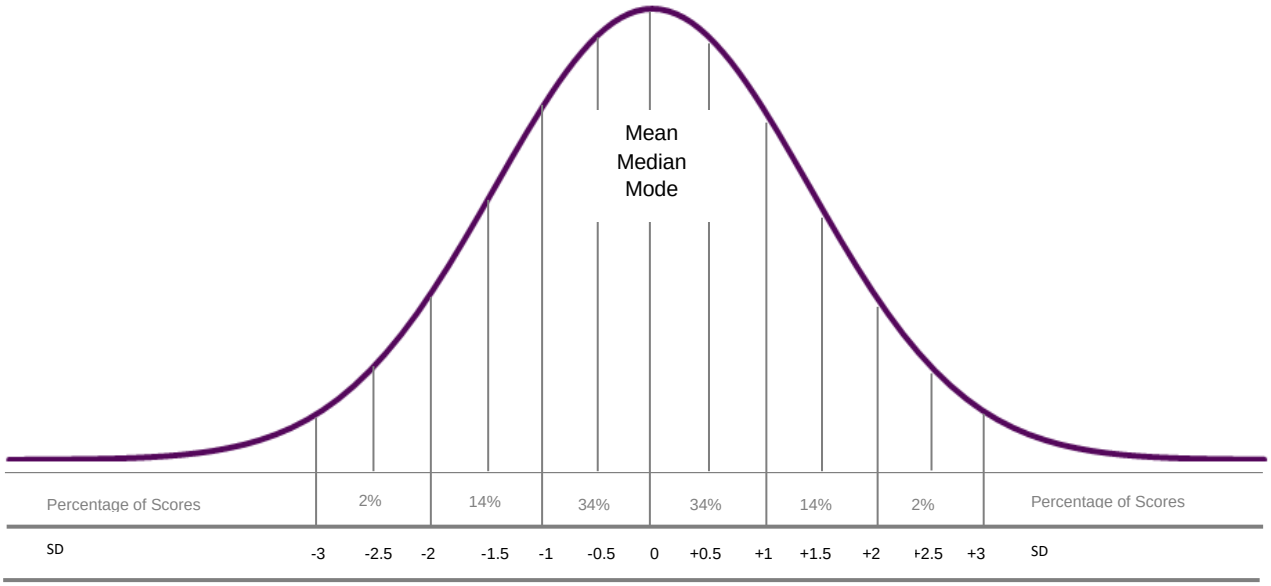
Under a normal distribution curve, the number of scores you would expect based on standard deviations from the mean are known. This information can help with the interpretation.

- 68.26% of observations will fall within one standard deviation either side of the mean
- 95.44% will fall within two standard deviations
- 99.74% will fall within three standard deviations

In practice these figures have been rounded up to make it a little simpler.

- 68% of scores fall within one standard deviation of the mean
- 96% of scores fall within two standard deviations of the mean
- 99% of scores fall within three standard deviations of the mean

Because of the very small number of observations falling above or below three standard deviations from the mean (less than 0.2 per cent), such extreme scores are rarely encountered in practice.

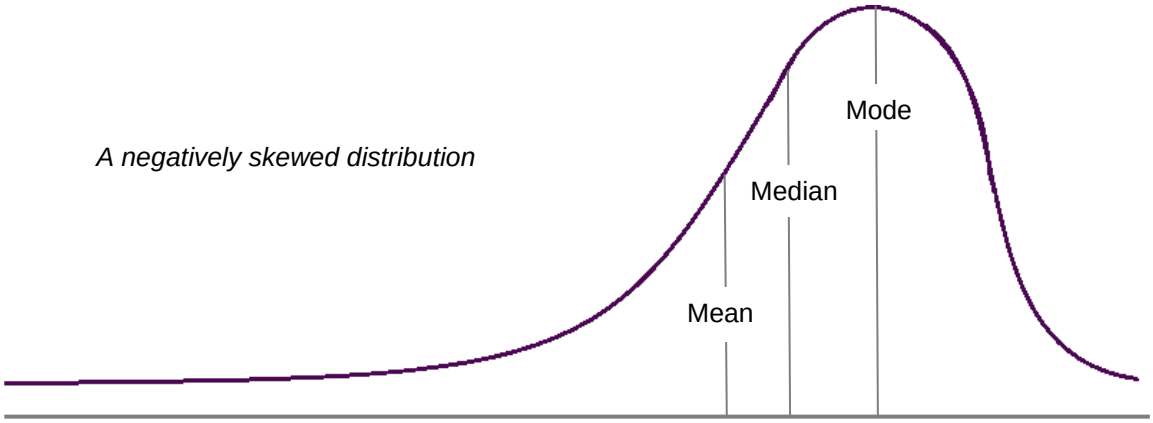


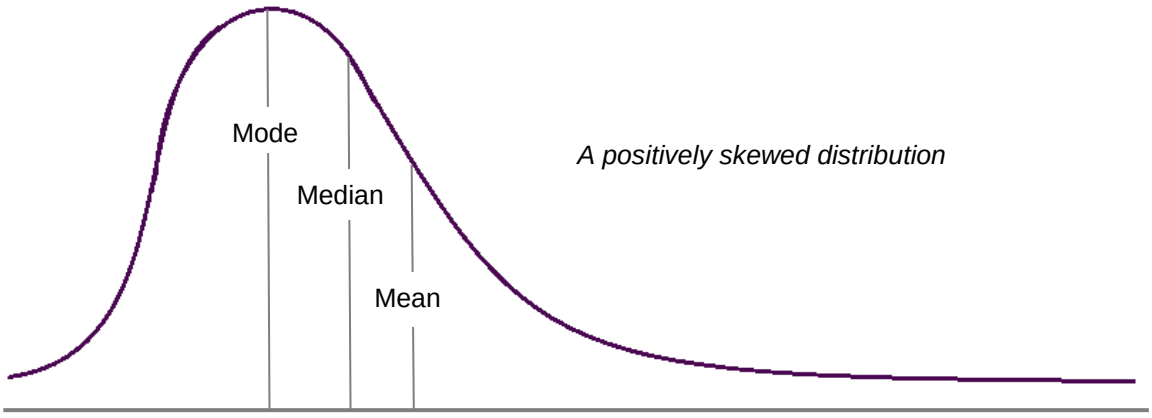
Skewed distributions

When using psychometric tests, sometimes the distribution of scores from certain samples of test-takers does not follow the normal distribution precisely. When a sample of test-takers' scores tend to score towards one end of the distribution, this is known as a 'skewed distribution'.

For example, a test of verbal reasoning ability may produce an approximately normal distribution when completed by a randomly selected group. The distribution of scores may, however, vary when the test is administered to other groups such as those with post-graduate qualifications or a sample from the general population. Test scores that tend towards the upper end of a score range are referred to as being 'negatively skewed', whereas those that fall towards the lower end of the score range are 'positively skewed', as illustrated below.

In a skewed distribution the mean, mode and median are no longer all at the mid-point as they are in a normal distribution. The highest point will always be the mode as it is the most common, but the median and the mean are pulled out towards the tail of the distribution. The diagrams below demonstrate the shift of the mean, median and modes. If the tail is pulled out towards the left this is known as a negative skew, as in the first diagram. If the tail is pulled out to the right it is referred to as a positive skew, as in the second diagram.





Interpreting test scores

When you describe the test scores it is important to understand what they mean. There are variations in the meaning or interpretations that are attached to test scores. There are three forms of referencing which affect the variations in interpretation.

Norm referencing

If a test-taker's score is compared to a group who have previously taken the test, then it is being referenced to a norm group. This is a widely used method for interpreting scores from psychometric tests. This method allows a person's test score (referred to as a 'raw score') to be interpreted relative to a specific comparison group, so judgements can be made as to whether it is below average, average or above average for the norm group.

An important point to recognise when using normative interpretations is that test scores are not fixed but will vary depending on what norm group is used. For example, a person may be above average when compared to a representative sample of school leavers, but only average when compared to a sample of generally more able graduates. When norm-referenced interpretation is used, scores are usually expressed either as a 'percentile' or a standardised score, both of which will be discussed later.

Domain referencing

For this form of referencing we require a clear definition of the construct the test is measuring and how this relates to the area we are interested in understanding. From this we make judgements about the level or amount of skill or knowledge a person has in the area assessed by the test.
One of the most common examples of this is school attainment exams, where the scores or grades obtained by test-takers indicate their level of knowledge in a given area.

Criterion referencing

This method uses test scores to predict performance on an outcome of interest to us. In criterion-referenced tests, the test items are often not directly related to what we are interested in predicting, as they are in domain referenced testing, but the abilities demonstrated through the test affect success on other outcome measures (e.g. job performance). For a criterion-referenced interpretation of a test score to be made, it is necessary for the link between test scores and the criterion we are interested in to have been established. This is done through a 'criterion-related validity study', which involves analysing the statistical relationship between test scores and the criterion of interest. Given the knowledge of this statistical relationship, it is possible to predict criterion performance from test scores.

Norm groups

When interpreting test results they can be compared to a large number of people who have previously completed the test. It is vital that a suitable group is selected when interpreting the scores; these groups are known as norm groups.

The norm group needs to be representative of the participant group. If used in selection the score for different groups would vary so their performance in comparison could look very different.

Norm groups can be selected from the test's user manual. Alternatively, organisations can create their own norm groups using previous applicant pools. According to the European Federation of Psychologists' Associations (EFPA) the minimum required number of participants is 150.

When selecting a norm group, this should be, where possible, a group who is comparable on a number of factors; age, gender, ethnic background and education. If it is not possible to match the norm group perfectly it is possible to select a group who have similar average scores to your test group. You must always aim to select the most representative norm group. This norm group should then be used on all candidates as separate norm groups could give different results. When using a test for development purposes, you could use separate norm groups, such as gender-specific norms as comparisons across candidates will not be made.

When developing a norm group, you are able to calculate the amount of error of estimation when using the performance of a sub-group to estimate the mean score of the total population. Any difference between the mean score for the sub group and the mean score for the total population is regarded as the error in estimation.

Using the 'standard error of the mean' formula below, we are able to see how large this error in estimation is likely to be.

The formula for the 'standard error of the mean':

$$SE_{mean} = \frac{SD}{\sqrt{N-1}}$$

Where:

SD = Standard Deviation

N = Number of people

A larger group will be more representative therefore the Standard Error of the Mean will be smaller.

Example

Charlie wanted to create a norm group for his organisation.

In his first attempt for a norm group he only had a sample of 50 people. He calculated the SEmean for this group.

For the first group the SD = 3 and N = 50.

$$SE_{mean} = \frac{SD}{\sqrt{N-1}}$$

$$\frac{3}{\sqrt{50-1}} = \frac{3}{7} = 0.43$$

The SEmean = 0.43.

In his second attempt he had managed to collect data for the recommended number of people (150). He calculated the SEmean for this group.

$$SE_{mean} = \frac{SD}{\sqrt{N-1}}$$

$$\frac{3}{\sqrt{150-1}} = \frac{3}{12.2} = 0.25$$

The SEmean = 0.25.

You can see from this example that as the number of people in the norm group increases the standard error of the mean decreases. In simple terms this means that as you use more people to act as a representation for the overall population you become more accurate at predicting what the average of the overall population would be. This enables you to trust your norm group more as the influence of individual differences are reduced.

Section 5 – Transforming Scores into Different Scales

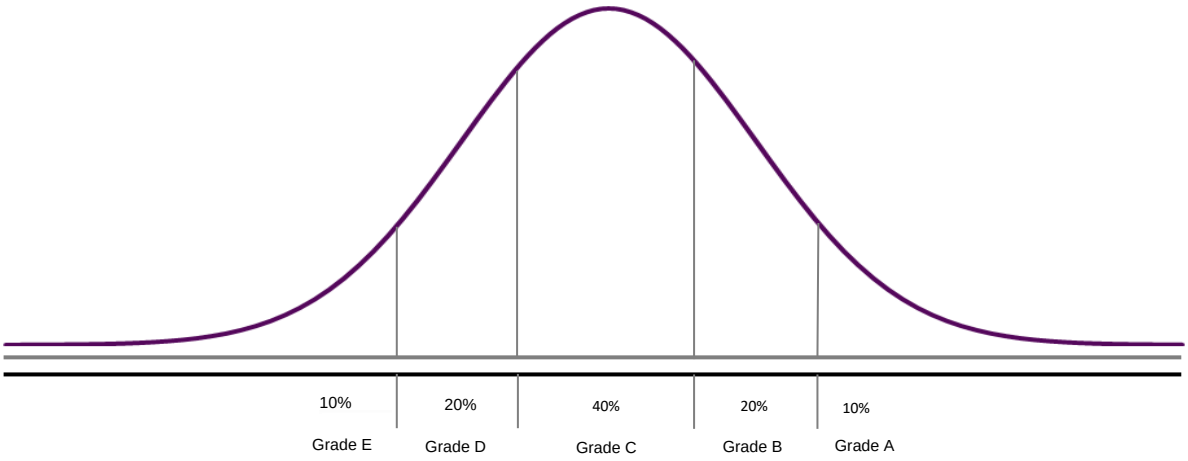
Introduction

In this section we will look at different scales to present test scores including grades, percentiles and standardised scales. We will look at the reasons for using these types of scores and ways of transforming test scores between different types of scales and how this information is presented in norm tables.

As most test scores and other measures of human attributes follow an approximately normal distribution, an understanding of the properties of the normal distribution helps us describe test scores in more meaningful ways.

Grades

Grades are the simplest norm system. They are used in educational testing and where broad indications of ability level are required. They are based on a five grade system. Grades on different tests cannot be added together as the raw scores are generated from different tests with different lengths.



Percentiles

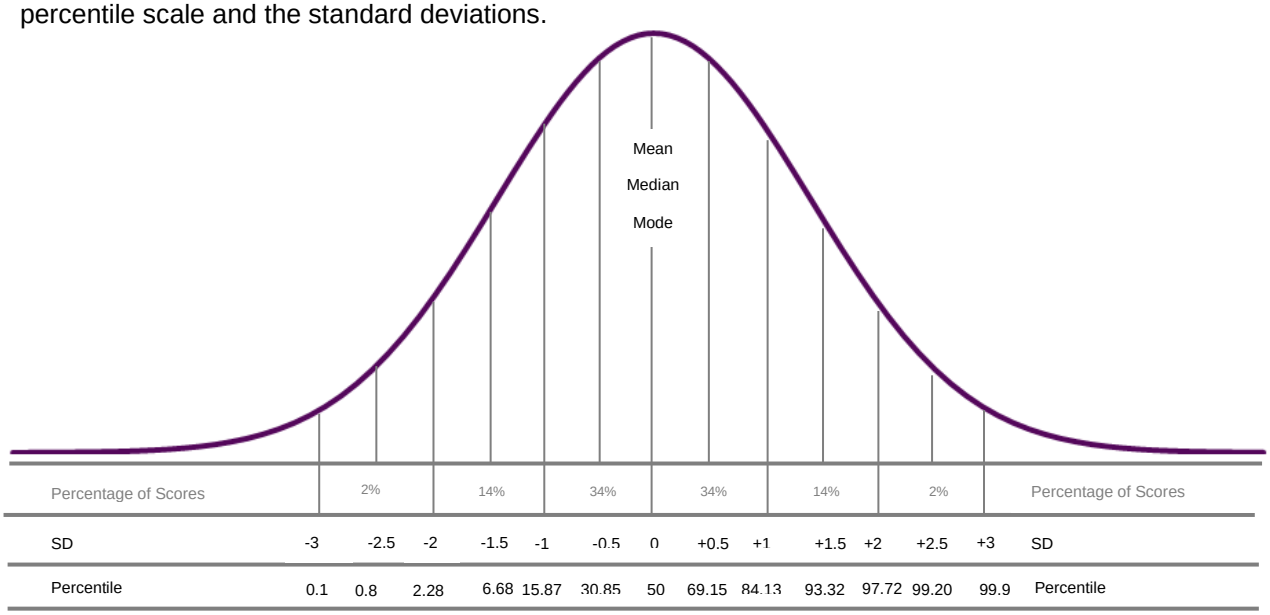
It is very common to want to compare an individual's score against the scores of the norm or comparison group. This approach to interpreting test scores produces values known as 'percentiles' or 'percentile ranks'.

A percentile rank represents the percentage of the norm group a raw score comes above.

For example, if a test-taker is said to have a percentile of 73, their raw score on the test is better than 73% of the comparison group. Alternatively, you could say they have scored within the top 27% of the comparison group.

Finding percentiles from the normal distribution

As you know the properties of a normal distribution, you can conclude that half of the scores will fall either side of the mean and that 68% of the score will fall within one standard deviation either side of the mean. Using this information, it is possible to calculate percentiles. For example; a score at the mean is equivalent to the 50th percentile. If the score is one standard deviation above the mean their score would be at the 84th percentile. The illustration below demonstrates the relationship between the percentile scale and the standard deviations.



Properties of percentiles

An important feature of the percentile distribution is that it is non-linear. There is a greater spread amongst scores towards the ends, or 'tails', and the distribution of scores in the middle are far more compact. Percentile scores are compacted towards the middle of the distribution as percentiles are a function of how many people score at each point in the distribution.

A second point to note is that because percentiles represent a ranking of candidates and are non-linear, they should not be used to perform any calculations. Therefore, if we wanted to sum a test-taker's scores from two tests to get an overall picture of their ability, we would have to use standardised scores in order to get a meaningful result.

Standardised scores

We can interpret scores by comparing them to a norm group or by using our understanding of the properties of the normal distribution. Standardised scores are a way of converting test scores into different scales, based on knowing the mean and standard deviation for a norm group.

There are a number of standardised scales that are used to describe test scores. These all describe test scores in a common way: the extent to which a test score varies from the mean in terms of standard deviation units.

Why do we need standardised scales?

The main reason for transforming the scores is because tests differ in their means and standard deviations, so it is impossible to compare raw scores from different tests. The means and standard deviations will vary due to tests containing different numbers of items, variations in the difficulties of the items, and variations in the samples used to compile norm groups.

In order to make meaningful comparisons between different tests, scores need to be on a common scale. This scale would need to take into account the mean and standard deviation of the individual test in order to allow a direct comparison between scores on different tests to be made.

Standardised scales can also be used for calculations because unlike percentiles which are non-linear and compacted towards the middles of the distribution, standardised scales are on a linear scale with equal intervals between values.

Z scores

Z scores describe the number of standard deviation units a score is above or below the mean of a norm group.

To transform a raw score into a Z score, it is necessary to know the mean of the raw scores and their standard deviation (SD). A raw score (X) is then transformed into a Z score using the following formula:

$$Z = \frac{X - \bar{X}}{SD}$$

Where:

Z = Standard Z score

X = Raw score

$\bar{X}$ = Mean score

SD = Standard deviation

Usually when standard scores are used they are interpreted in relation to the normal distribution curve. On the normal distribution curve you can see that the standard deviation units are marked out either side of the mean. Those above the mean are positive, and below the mean are negative in sign. Therefore, by calculation of the standard Z score it can be seen where the individual's score lies in relation to the rest of the distribution.

For example:

Raw Score (X) = 30

Mean ($\bar{X}$) = 20

Standard Deviation of a Test (SD) = 5

$$Z = \frac{X - \bar{X}}{SD} = \frac{30 - 20}{5}$$

$$= \frac{10}{5}$$

$$= 2$$

The raw score of 30 is 2 standard deviations above the mean. This score is equivalent to a percentile of approximately 98.

Because standard scores are equal units of measurement they can be mathematically manipulated and therefore have an advantage over grades and percentiles.

The standard score is very important when comparing scores from different tests. Before these scores can be properly compared they must be converted to a common scale such as the standard Z score. One important advantage in using the normal distribution as a basis for test norms is that the standard deviation has a precise relationship with the area under the curve. One standard deviation above and below the mean includes approximately 68% of the sample.

Standard **Z scores have a mean of 0** and **a standard deviation of 1**. Because of this, Z scores can be rather difficult to handle as most of them are decimals and approximately half of them can be expected to be negative. To remedy these drawbacks various transformed standard score systems have been derived. These simply entail multiplying the obtained Z score by a new standard deviation and adding it to a new mean. Both of these steps are devised to eradicate decimals and negative numbers.

T scores

The T score is the most common standardised norm system for ability tests. The T score is derived from the Z score which has been transformed to a new scale, so that the **T score mean is set at 50** and the **standard deviation at 10.**

To calculate the T score:

T score	= (Z score x 10) + 50
If Z score	= 2
T score	= (2 x 10) + 50
	= 20 + 50
	= 70

The advantage of T scores over percentiles is that they are a linear scale with equal units of measurement. Thus they can be mathematically manipulated and give results which can be directly compared both within and between individuals.

Their disadvantage is that they are not easily explained to those not knowledgeable about testing.

Sten

Sten is an abbreviation for 'Standard Ten', and is a standard score system which divides the score scale into 10 divisions. It is based on a **mean of 5.5** and a **standard deviation of 2.**

Sten	= (Z score x 2) + 5.5
If Z score	= 2
Sten	= (2 x 2) + 5.5
	= 9.5

Normally, the sten is taken to the nearest whole number, with a minimum value of 1 and a maximum of 10. Stens are commonly used as the norm system for personality questionnaires. However, they are also a useful system for ability tests as they encourage us to think in bands of scores but still provide sufficient discrimination between applicants.

Stanine

Stanine is an abbreviation of 'Standard Nine', and is similar to the sten score above, but dividing the score scale into 9 divisions. It is based on a **mean of 5** and a **standard deviation of 2.**

Stanine	= (Z score x 2) + 5
If Z score	= 1.5
Stanine	= (2 x 1.5) + 5
	= 8

Standard Score IQ

The standard score IQ is a further norm system. IQ stands for 'Intelligence Quotient' and the scale is commonly used with tests designed to measure intelligence.

As a scale, its **mean is set at 100.** The **SD is 15,** and based on this the IQ is calculated as:

IQ	= (Z score x 15) +
100	
If Z score	= 2
IQ	= 2 x 15 + 100
	= 30 + 100
	= 130

Besides being used with intelligence type tests, IQs are a common norm system for use with attainment and ability tests used in the educational sphere, and more recently with tests of Emotional Intelligence.

Mean
Median
Mode
Percentage of Scores
2%
14%
34%
34%
14%
2%
Standard (Z) Score
-3
-2.5
-2
-1.5
-1
-0.5
0
+0.5
+1
+1.5
+2
+2.5
+3
T Score
20
25
30
35
40
45
50
55
60
65
70
75
80
Percentile
0.1
0.8
2.28
6.68
15.87
30.85
50
69.15
84.13
93.32
97.72
99.20
99.9
IQ
55
62.5
70
77.5
85
92.5
100
107.5
115
122.5
130
137.5
145
Sten
1
2
3
4
5
6
7
8
9
10
Stanine
1
2
3
4
5
6
7
8
9

The normal distribution curve illustrates where the Z scores are, by using the distributions of the proportion of people falling in different areas we can interpret the Z score. We are able to see that approximately 16% of the distribution falls to the left of the Z score -1, making a Z score of -1 equivalent to a percentile of 16. Our Z score of 0 is at the middle of the distribution, making it equivalent to a percentile of 50.

There is no formula that allows us to convert from Z scores into percentiles and vice versa. The way we find the exact percentile equivalent of a Z score is by using a conversion table.

Standard score and percentile conversion table

Z Score	T Score	Percentile
-3.0	20	0.10
-2.8	22	0.20
-2.6	24	0.40
-2.4	26	0.80
-2.2	28	1.50
-2.0	30	2.30
-1.8	32	3.50
-1.6	34	5.50
-1.4	36	8.00
-1.2	38	11.50
-1.0	40	16.00
-0.8	42	21.00
-0.6	44	27.50
-0.4	46	34.50
-0.2	48	42.00
0.0	50	50.00
+0.2	52	58.00
+0.4	54	65.50
+0.6	56	72.50
+0.8	58	79.00
+1.0	60	84.00
+1.2	62	88.50
+1.4	64	92.00
+1.6	66	94.50
+1.8	68	96.50
+2.0	70	97.70
+2.2	72	98.50
+2.4	74	99.20
+2.6	76	99.60
+2.8	78	99.80
+3.0	80	99.90

Norm tables

The main purpose of using percentiles and standardised scales is to simplify the interpretation of test scores. We have covered how to calculate percentiles and standardised scores, and how to transform scores between different standardised scales. This understanding is important when looking at the reliability and standard errors, and when making comparisons between the scores of different test-takers. When initially interpreting test scores, however, the test's norm table is one of the most useful references.

Norm tables provide a way of transforming raw scores into percentiles and standardised scores. They will have been developed by the test publisher and so it is unlikely that there is a need to do any calculations.

In the following example of a norm table you are able to see that there is one column showing the possible raw scores that can be obtained on each of the four tests as well as the corresponding Grade, Percentile, T-score or Sten Score.

A norm table is a convenient way of looking up raw scores and translating them into other types of score. It is important to remember, however, that interpretations made this way are being made relative to the norm group. This means that scores are not fixed, but are dependent on the norm group used: a score may be high in relation to one group but only average in relation to another, generally more capable group.

Example norm table for a group of tests

		Test A	Test B	Test C	Test D		
Grade	**%ile**					**T Score**	**Sten**
A	99			35-36		75	10
	99			34	36	74	
	99			33	35	73	
	99					72	
	98			32	34	71	
	98			31	33	70	
	97				32	69	9
	96	36	40	30	31	68	
	96	35	39	29		67	
	95			28	30	66	
	93	34	38		29	65	
	92	33	37	27	28	64	8
B	90	32	36	26		63	
	88		35		27	62	
	86	31	34	25	26	61	
	84	30	33	24	25	60	
	82	29	32		24	59	7
	79		31	23		58	
	76	28	30	22	23	57	
	73	27	29	21	22	56	
C	69	26	28		21	55	6
	66			20		54	
	62	25	27	19	20	53	
	58	24	26		19	52	
	54	23	25	18	18	51	
	50		24	17	17	50	
	46	22	23			49	5
	42	21	22	16	16	48	
	38		21	15	15	47	
	34	20	20	14	14	46	
	31	19	19			45	
D	27	18	18	13	13	44	4
	24			12	12	43	
	21	17	17		11	42	
	18	16	16	11	10	41	
	16	15	15	10		40	
	14		14		9	39	3
	12	14	13	9	8	38	
	10	13	12	8	7	37	
E	8	12	11	7	6	36	2
	7		10			35	
	5	11	9	6	5	34	
	4	10	8	5	4	33	
	4	9			3	32	
	3		7	4		31	
	2	8	6	3	2	30	1
	2	7	5		1	29	
	1	6	4	2	0	28	
	1		3	1		27	
	1	5	2	0		26	
	1	0-4	0-1			25	

Section 6 – Measuring Reliability & Validity

Introduction

As we have already established it is important that tests are both reliable and valid. Calculations and statistical information can be used to identify whether the tests meet the required standards. This section will guide you through the statistics which are used and demonstrate how you can calculate the values of a test's validity and reliability.

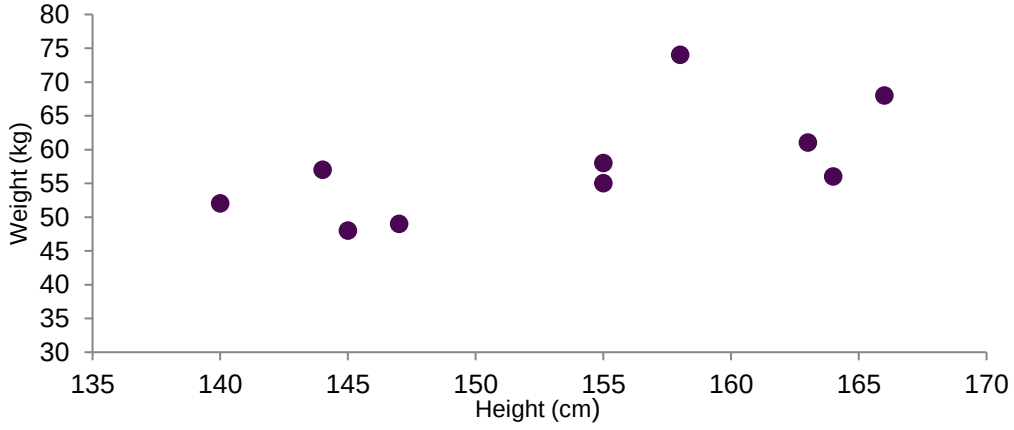
Correlations

A correlation is a statistical technique widely used in understanding reliability and validity. It looks at the relationship between any two variables.

Example:
The height and weight data for ten people randomly selected from the general population.

Person	Height (cm)	Weight (kg)
Ellie	140	52
James	144	57
Amy	145	48
Katie	147	49
Polly	155	55
Tamsin	155	58
Brad	158	74
Dylan	163	61
Andy	164	56
Lucia	166	68

By looking at this table you can see that there is a general trend, with taller people being heavier although this is not the case for everyone. This data can be displayed on a scatter-plot as illustrated below:

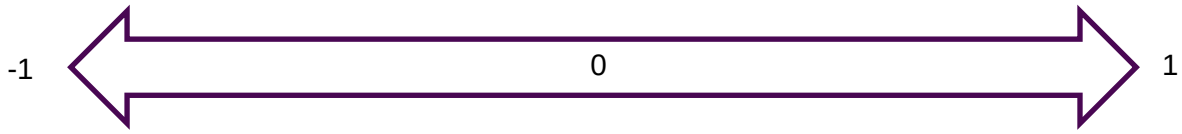


The scatterplot confirms the general trend with the pattern moving from the bottom left corner up to the top right, showing that people who are taller tend to be heavier.

Tables and scatterplots are useful ways of presenting the data but make it difficult to accurately say how strong the relationship is. By just looking at the scatterplot can you confidently say that there is a strong relationship between the two variables? Correlations can help with this as they summarise the data in the table to produce a single statistic indicating the strength and direction of the association between the two variables we have measured.

Correlation coefficients

Calculating the association between two variables using correlations will produce a value between -1 and +1. This value is referred to as the 'correlation coefficient' and is often indicated by the letter 'r'. The most commonly used formula for calculating the correlation coefficient is the Pearson Product Moment Correlation formula.

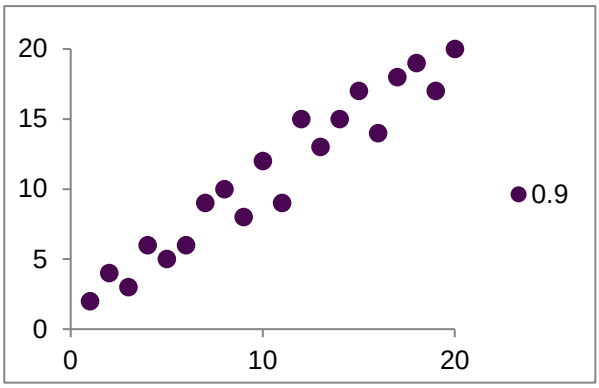


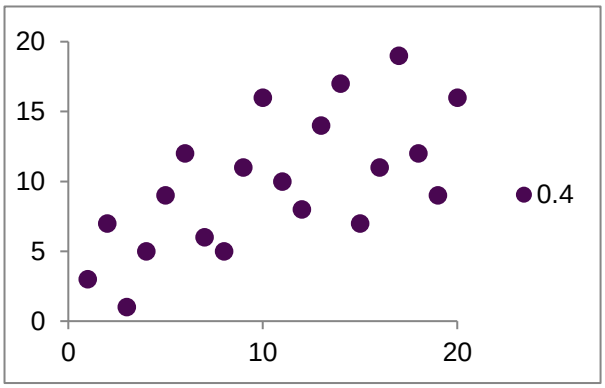
The **sign** of the correlation (positive or negative) indicates the **direction** of the association. In a positive correlation when one variable increases so does the other, whereas a negative correlation indicates that as one variable increases the other decreases.

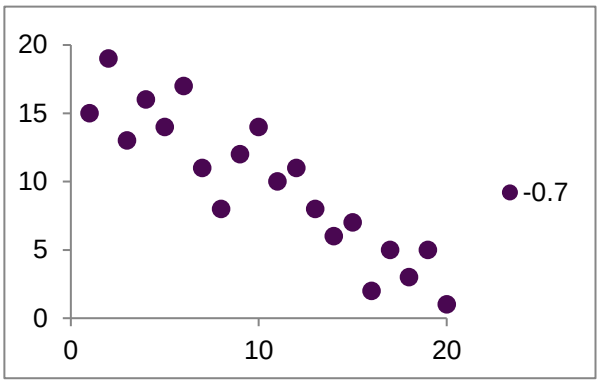
The **number** given to the coefficient indicates the **strength** of the association. Correlation coefficients closer to 1 (whether it is positive or negative) indicate that there is a stronger association between the two variables. Coefficients closer to 0 indicate a weaker association between the variables.

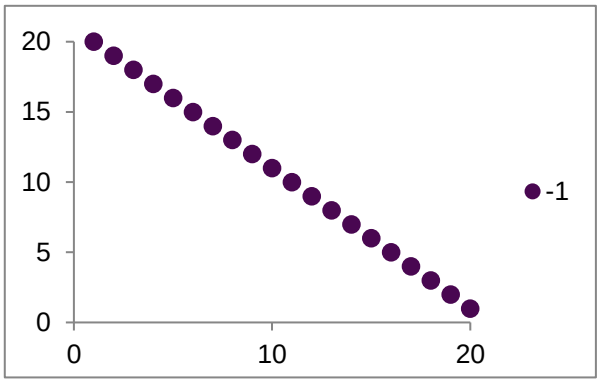
There are no set rules for interpreting the strength of a correlation coefficient as it is dependent on the context whether it is classed as 'strong' or 'weak'. It is also dependent on the context and type of data to which the correlation is being applied.

Here are some examples of correlation coefficients of different strengths and directions and their corresponding scatter-plots. There is a much wider spread of scores for those with weaker correlations. The spread of scores on a scatterplot with a strong correlation is much more unified.









Formula for Pearson's Product Moment Correlation Coefficient

The correlation between two variables can be calculated using the formula below.

Pearson's Product Moment Correlation Coefficient:

$$r = \frac{\Sigma\, xy}{SD_x\, SD_y N}$$

Where:

r = the product moment correlation coefficient
Σ = the sum of
x = the deviation of a person's raw score on variable 1 from the mean
y = the deviation of a person's raw score on variable 2 from the mean
SD_x = the standard deviation on variable 1
SD_y = the standard deviation on variable 2
N = the number in the group

In this formula the statistical notations 'x' and 'y' have been used. These are the shorthand expressions for the deviation scores (X – $\bar{X}$) and (Y - $\bar{Y}$) respectively.

$x = X - \bar{X}$

$y = Y - \bar{Y}$

On the next page is a worked example which takes you through each of the stages.

Example

A group of six people are given a numerical reasoning test, and these are correlated with a measure of their job performance.

	Scores on Test (X)	Job Perf. Scores (Y)
Chris	30	66
Karen	28	75
Gary	32	70
Daisy	27	65
Mike	30	74
Gillian	33	70

Using the Mean for the test scores and job performance it is possible to calculate the deviation on each of these.

$\bar{X}$ = 30 $\bar{Y}$ = 70

SD_x = 2.08 SD_y = 3.70

N = 6

	Scores on Test (X)	Job Perf. Scores (Y)	Deviation Score on X (x)	Deviation Score on Y (y)	Cross Product of Deviation Scores (xy)
			$x = X - \bar{X}$	$y = Y - \bar{Y}$	Deviation on X multiplied by Deviation on Y
Chris	30	66	30 – 30 = 0	66 – 70 = -4	0 x -4 = 0
Mike	28	75	28 – 30 = -2	75 – 70 = 5	-2 x 5 = -10
Gary	32	70	32 – 30 = 2	70 – 70 = 0	2 x 0 = 0
Daisy	27	65	27 – 30 = -3	65 – 70 = -5	-3 x -5 = 15
Karen	33	74	33 – 30 = 3	74 – 70 = 4	-3 x -4 = 12
Gillian	30	70	30 – 30 = 0	70 – 70 = 0	0 x 0 = 0

Find the sum of the cross product of deviation of scores.

$\sum xy$ = 0 + -10 + 0 + 15 + 12 + 0 = 17

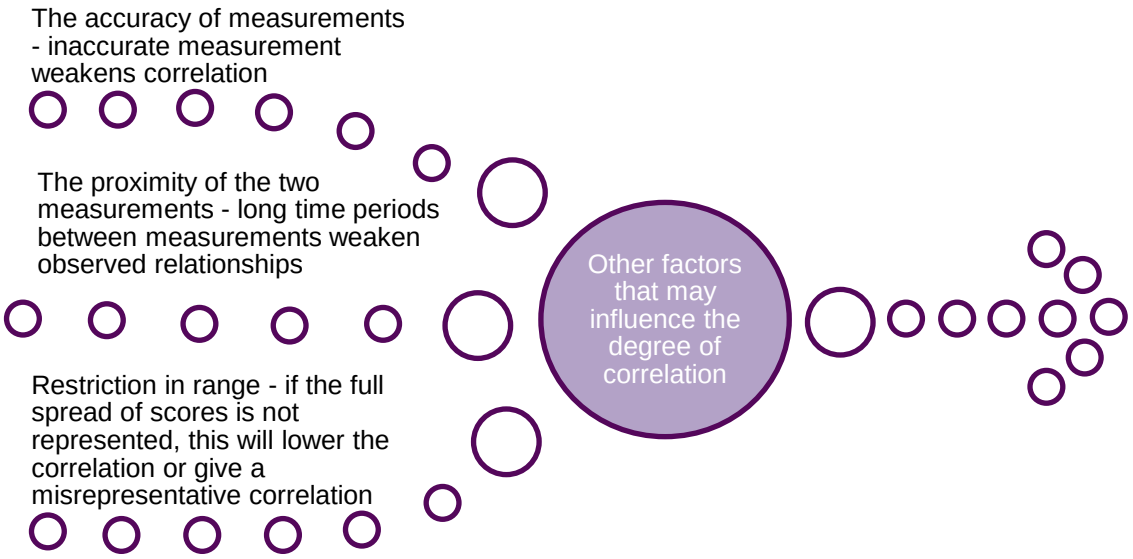
Then substitute the values into the formula:

$$r = \frac{17}{6 \; x \; 2.08 \; x \; 3.70} = \frac{17}{46.18} = \mathbf{0.37}$$

Factors affecting correlation coefficients

There are many factors which may affect the degree of relationship seen between two variables and it is important to consider these when interpreting correlation coefficients. If the underlying relationship between the two variables is strong, then this should show itself through a stronger correlation coefficient (whether it is positive or negative).

Other factors influencing the degree of correlation observed include:

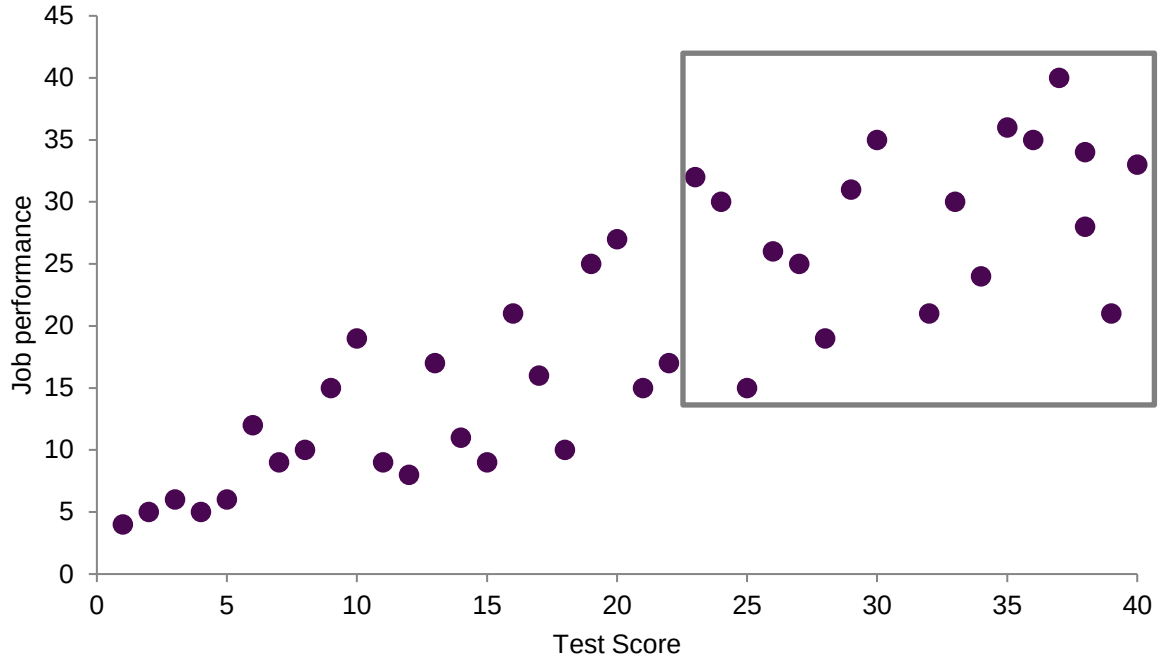


Restriction of range and correlation coefficients

Correlations are used a lot in psychometrics, especially for calculating the reliability and validity of tests which will be explained in more depth later in this section. However, they can be affected by a 'restriction of range'; this is when a full spread of scores is not available for one or both of the variables being correlated. If there is a restriction in range, the observed correlation will be lowered, even if the true relationship between the variables is a strong one.

There are many reasons why a restriction in range may occur, but it is most common in selection contexts when the test scores are used to aid the decision making process for recruitment. If a test user wants to correlate test scores with subsequent job performance part of the sample will not be included in the analysis because they were not selected. Importantly, the proportion of the sample that are lost to the analysis will not be random as typically these will be the lower performing applicants.

Restriction of Range

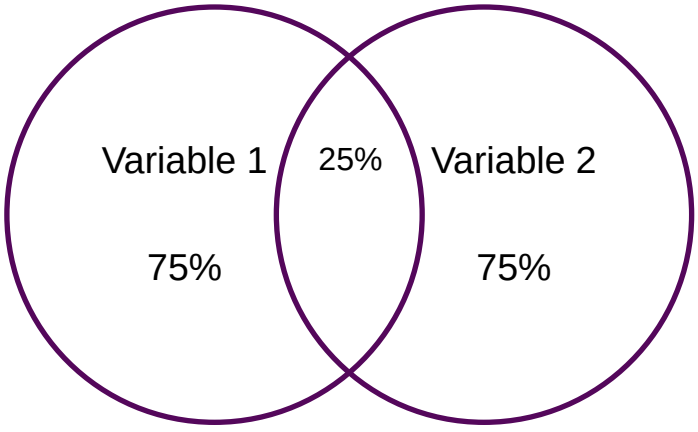


If the range is restricted to just those candidates in the square box, then clearly the strength of the correlation is much weaker. The spread is much wider within the square box compared to the trend which can be seen if you take all of the data points into account.

Correlation and shared variance

A correlation coefficient indicates the degree of association between two variables. Another way to interpret the correlation is to think of it in terms of the shared variance or overlap between the two variables, which can be calculated by squaring the correlation coefficient.

For example, if the correlation coefficient between two variables came to 0.5, this could be squared to give 0.25 indicating that the two variables had 25% of their variance in common.



The illustration demonstrates that the two variables have 25% of their variance in common, with each having 75% that is independent of the other.

We will come back to shared variance later, as it is useful when exploring how much of the variation in areas such as job performance can be explained by tests and other assessment methods.

Correlation and causality

When looking at the relationship between variables by using correlations it can be tempting to conclude that the change observed in one of the variables is the cause of the changes in the other. Whilst this may be the case, the correlation itself does not tell us this; it merely shows whether there is a link between the two variables.

Consider the following example:

> There is a positive correlation between ice cream sales and the amount of sun-tanning experienced by a group of people
>
> Therefore:
>
> Eating ice cream increases the likelihood of sun-tanning.
>
> In fact…
>
> The correlation between ice cream and sun tanning is down to a third factor: heat!
>
> When it is hotter ice cream sales increase and more people sun bathe, giving rise to the increased likelihood of sun-tanning.

In practice, we may be reasonably confident that a link in a certain direction does exist between two variables. For example, if we continue with the scenario above, we would be fairly confident that more people sun-bathe in hotter weather, and not that more people sun-bathing causes an increase in temperature. This conclusion, however, is going beyond what the correlation coefficient itself actually tells us by providing an additional interpretation to the data.

Measuring reliability

The value given to represent a test's reliability is a correlation coefficient. There are several different methods for calculating the reliability of a test and these all vary depending on which factors are correlated.

- **Internal consistency**: scores on odd questions vs. scores on even questions
- **Test – retest reliability**: scores at time 1 vs. scores at time 2
- **Parallel form**: scores on test 1 vs. scores on test 2

Each of the methods for estimating the reliability of a test produces a value which indicates the accuracy or reliability of the test. This value will range from 0 to 1, with values closer to 1 indicating higher reliabilities.

Standards around reliability are described by the British Psychological Society (BPS), along with guidelines for interpreting reliability values. For a test to be sufficiently accurate, its reliability value should be **at least 0.7**. It is also important that any reliability figure quoted for a test is based on an adequate number of respondents, as larger sample sizes will give more robust results. Here the BPS guidance is that the samples used in reliability studies should include at least 100 people.

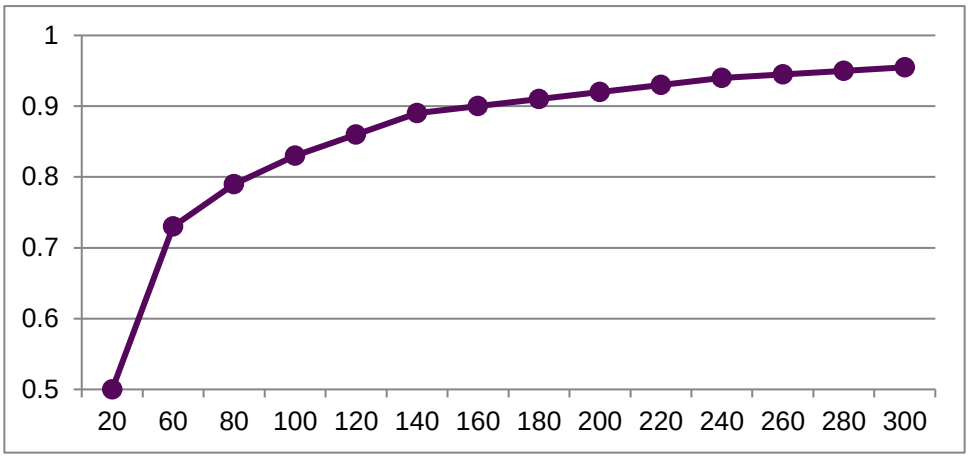
When studying the reliability values given in test manuals, it will be apparent that they vary quite considerably between different types of tests. As a rule, tests assessing broader, more general abilities such as verbal critical reasoning will tend to have lower reliabilities than those assessing more homogenous, clearly defined constructs such as numerical ability. Tests that are highly speeded (i.e. where performance is significantly determined by how fast a test-taker works at the test) also tend to show higher internal consistency reliabilities than tests where time is not such a factor. Some consider this speed element to artificially inflate the reliability estimate, and so recommend seeking evidence of test-retest or parallel from reliability in such cases.

Reliability and test length

One of the main factors affecting the reliability of a test is its length, with longer tests, on average, being more reliable than shorter ones. Any single test item may contain a large amount of error, but as the number of items in a test increases the proportion of measurement error decreases.

Intuitively the relationship between test length and reliability makes sense to test users. No matter how efficient in terms of time, more test users would be more confident in the results from a twenty-five item test than the results from a five item one.

The relationship between test length and reliability is not linear, meaning that doubling the number of items will not result in doubling of a test's reliability. The curvilinear relationship between reliability and test length is shown overleaf.



Recognising the error in test scores

Earlier we saw that a number of factors can influence the reliability or accuracy of the measurements that are obtained from a test. Selecting well-developed tests, preparing test-takers and standardised administration will all help to increase reliability, but some error in measurement will always remain. The test publishers will report the reliability values in the test manuals after investigating the reliability during the development process. If we use tests in a standardised manner, we can be confident that the reliability values in the manual apply to our particular applications of the tests due to the theory of generalisability.

One of the advantages of psychometric tests over other forms of assessment is that the error in psychometric tests is made explicit. Developers and users of tests acknowledge that test measurement is not perfect and results should be treated as being an indicator of a person's standing on the construct being measured.

The Standard Error of Measurement

In the same way that we use statistics to describe the distribution of observed scores from a test, we can also use statistics to describe the distribution of scores due to error. The statistic that we use to do this is the Standard Error of Measurement (SEm). For any observed score, the SEm allows us to identify the range or score band within which a test-taker's true score is likely to be, with a given degree of confidence.

This score band can be calculated by working out the standard error of measurement using the formula below and then accepting a score between 1 unit of SEm above and below the observed score.

$$SEm = SD\sqrt{(1 - r)}$$

*Where SD is the standard deviation of the test and *r* is the reliability of the test

Calculation of the Standard Error of Measurement

A total of 68 people were given a Numerical Reasoning Test. Their scores on the test had a SD of 6.2 and a reliability coefficient of 0.89. The SEm calculation is below:

$SEm = 6.2\sqrt{(1 - 0.89)}$

SEm = 2.0563

SEm = 2.06 *2 decimal places

Interpreting the SEm

The standard error of measurement (SEm) can be defined as the standard deviation of scores due to error. There are two assumptions:

- The obtained raw score for an individual is the best estimate of that person's ability.
- Variations from this best estimate tend to be normally distributed.

On these assumptions from our knowledge of the normal distribution curve we can say that:

- There is a **68% probability** that a person's ability lies within one unit of SEm (±1 SEm) of the score obtained.

- There is a **95% probability** that a person's ability lies within two units of SEm (±2 SEm) of the score obtained.

Adding the SEm to either side of the observed score provides you with a confidence band within which the true ability will fall.

For a Numerical Reasoning Test with an SEm of 2, if an applicant scored 22:

- there is a 68% probability that his 'true' ability would fall between 20 and 24
- there is a 95% probability that his 'true' ability would fall between 18 and 26.

Differences between two applicants

There is the danger that we might place too much emphasis on small differences between candidates. The next example demonstrates how important the confidence interval is in distinguishing between results within a group.

Example: Six graduates have completed a test.

Name	Score
Clifford	35
Smith	27
Peters	24
Carter	22
Williams	16
Larson	15

Assume SEm = 2.5

Can we be sure that there is a real difference between the candidates?

As a general guide, you should allow a difference between two applicants of two units of SEm to have sufficient confidence that there is a real difference between them.

For example, the difference between Clifford and Smith is 8. This is greater than 2 units of SEm (2.5 x 2 = 5) therefore Clifford and Smith are different. However, the differences between Smith, Peters and Carter are probably due to chance.

To calculate whether a person has performed better on one test than on another we can calculate the **standard error for the difference** between the two scores. This is based upon the standard errors of measurement of the two tests and is given by this formula:

Standard Error for the Difference

$$SEd = \sqrt{(SEm_1{}^2 + SEm_2{}^2)}$$

Where SEd = Standard error of a difference between two scores for the same individual on two tests
SEm_1 = Standard error of measurement for the first test in standard score form
SEm_2 = Standard error of measurement for the second test in standard score form

Example:

If one test had a SEm of 4.6 T scores and another had 3.6 T scores, then the Standard Error of Difference would be:

$$SEd = \sqrt{(4.6^2 + 3.9^2)}$$
$$= \sqrt{36.37}$$
$$= 6.03$$

If the difference between scores on two different tests for an individual is greater than 6.03, we can deduce that there is a significant difference between them.

This equation can also be used when calculating the difference between candidates who complete the same test. This would tell you whether someone is significantly better than another person on a test.

Example:

If the test had a SEm of 4.6 T scores we would calculate the Standard Error of Difference by replacing the second SEm with the SEm of the test again. The formula:

$$SEd = \sqrt{(SEM_1{}^2 + SEM_1{}^2)}$$
$$SEd = \sqrt{(4.6^2 + 4.6^2)}$$
$$= \sqrt{42.32}$$
$$= 6.51$$

Therefore, when comparing two candidates on the test they would need to have a difference of 6.51 between their scores to be regarded as different.

Use of reliability data

A reputable test always provides information on reliability. If no reliability data is given in the manual it is best not to use the test.

Below is an example of the reliability information which would be provided for a test battery. This shows internal consistency reliability coefficients, based on Cronbach's Coefficient Alpha. The reliability coefficients range from 0.86 to 0.92. Standard errors of measurement (SEm) are also given.

Internal consistency reliability data from the Saville Consulting Ability Test manual

Test	N	$\overline{X}$	SD	Alpha r	Raw Score SEm	T-Score SEm
Commercial Verbal Comprehension	164	16.13	4.21	0.75	2.11	5.01
Commercial Numerical Comprehension	164	14.41	5.09	0.85	1.97	3.87
Commercial Error Checking	164	13.15	4.39	0.82	1.86	4.24

Generalisability

We have already looked at where error can come from. If we are going to use the test in a setting where the sources of error are going to be roughly the same as the study where the reliability estimate was calculated, then it is reasonable to take the reliability estimate as applicable to our situation.
If we use the test in more controlled conditions the sources of error would be reduced therefore the reliability estimate would underestimate the accuracy of the measures.

An alternative to this classical approach to reliability is the **generalisability theory**. This theory measures the effect of each separate source of error which will be present. Using this information, it is possible to estimate the amount of error in the measures obtained from various different sets of conditions. The name comes from the fact that this approach is concerned with how much we can generalise across conditions. This approach can be difficult in practice to establish a data-base to generalise from.
As most occupational psychology testing tends to be carried out across situations with comparable error sources, following rigorous procedures for standardisation, the classical approach is appropriate despite being a less sophisticated method.

The different methods of generating reliability measures can give different results for the same test as there are different levels of control over conditions.

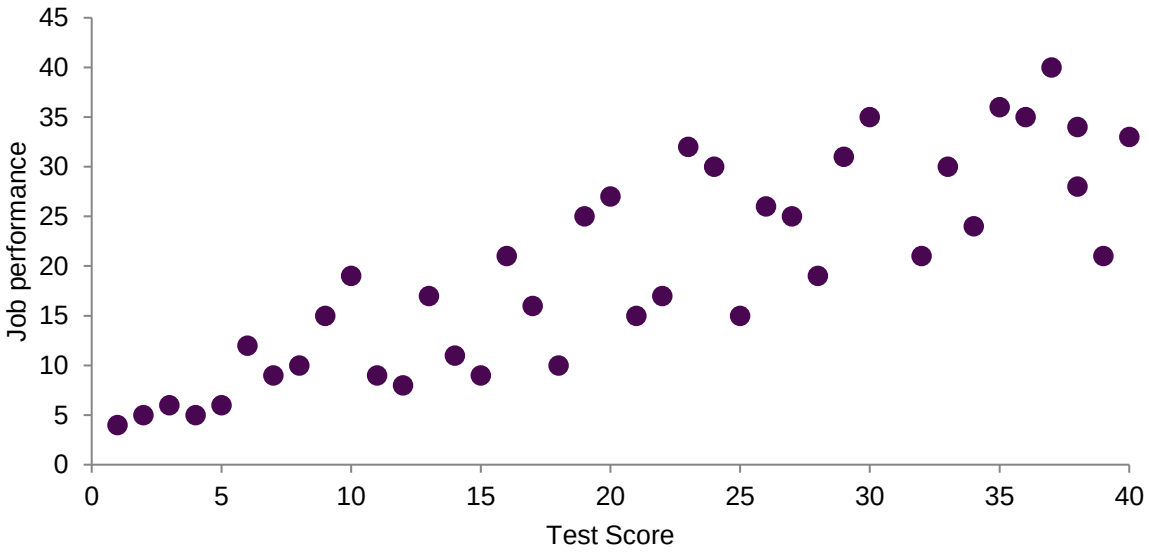
- Test-retest correlations tend to be higher than parallel form correlations as item variation is removed so there is one less source of error.
- Internal consistency also tends to be higher than parallel form correlations because the latter introduce the time interval between tests which is a source of error.

By reducing the range of true scores in a sample there would be a drop in the reliability as the relative proportion of error in the total variance increases.

Measuring Validity

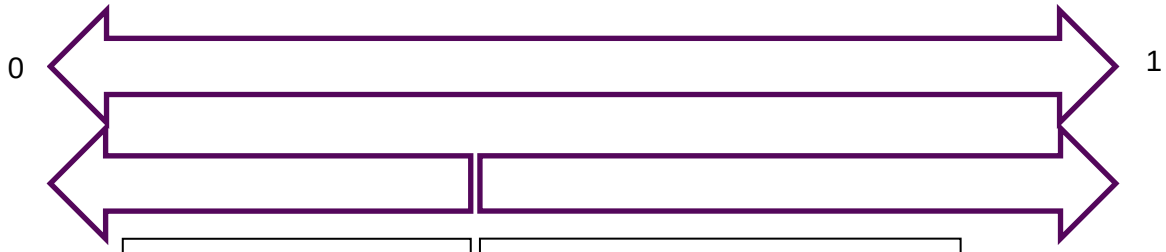
Validity is measured using correlations to investigate the association between the test scores and aspects of job performance.

We discussed the types of validity in the section introducing validity and identified that criterion validity relies on the association between test scores and criteria measures. To establish whether the association exists we would need to take both test scores and criteria measures from a sample of employees. We can do this in one of two ways: by taking test scores and criteria measures both at the same time, or by taking test scores at one point in time and criteria measures sometime later. Each employee's score on both methods can then be plotted on a scatterplot like the one below.



Validity coefficients

The value for the validity of the test is the correlation coefficient produced when the performance scores are compared with the test scores. The 'r' value is then used to represent the validity for the test for that purpose. As it is a correlation coefficient the value will be between 0 and 1. Validity studies are usually conducted with a large sample and a typical level of validity for ability tests is around 0.4 to 0.5. Tests with validity values below 0.3 are unlikely to be useful unless under exceptional conditions.

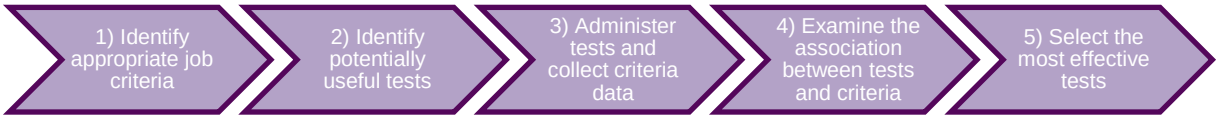


Validity values of 0.3 or less are unlikely to be useful unless under exceptional conditions	The range of acceptable validity is usually 0.4 or above. A typical level of validity for ability tests is 0.4 to 0.5

Concurrent Validity

This is where the tests and criteria are measured at the same point in time. It can be useful to give an initial understanding of whether certain tests would be good indicators of job performance. However, concurrent validity does not guarantee that a test would work in a predictive way. The results of this kind of study may be affected by several factors. Learning, experience or job-related training may impact on both job performance and test scores. The motivation of test-takers may also affect scores.

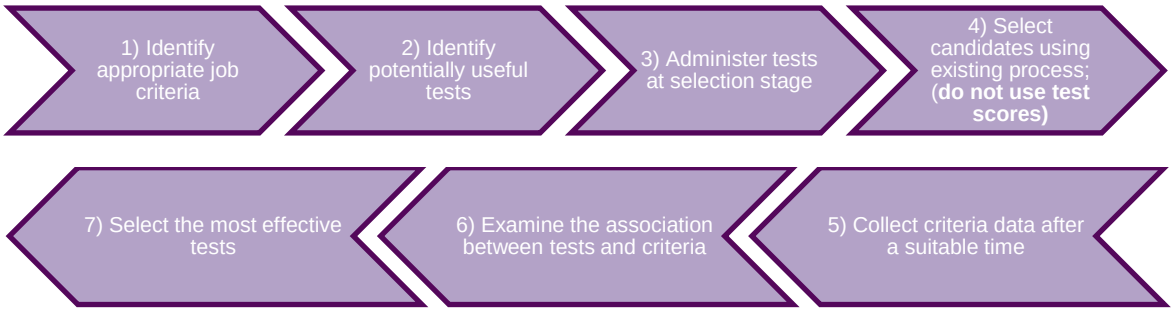
The stages involved in a concurrent validity study are:



Predictive validity

This type of a validity study replicates the selection process more accurately as it investigates how applicants will perform on the job at some point in the future. For this reason, predictive validity provides some of the strongest evidence of a test's validity. However, it does have its limitations as it can take many months to gather the criterion data.

The stages involved in a predictive validity study are:



Note that stage four says that tests scores should not be used as part of the selection process at this stage. This is because the effectiveness or otherwise of the test has not been established at this point. Using the test scores may result in selection decisions which may be less accurate than using existing methods alone.

The table summarises the benefits and issues with concurrent and predictive validity studies.

Concurrent Validity Studies	**Predictive Validity Studies**
Quick to conduct as test scores and criteria are measured at same time	Takes time to conduct as test scores and criteria are measured at different times
Does not accurately reflect the employee selection process	Accurately reflects the employee selection process
Useful for refining an initial selection of tests to use in a predictive study	Useful for identifying tests that can be used in the actual selection process
Motivation of test-takers may be different from applicants	Job applicants are likely to be highly motivated to complete tests
Results can be affected by experience, training etc.	Results are less affected by experience and training
Relatively cheap to conduct	Typically, costlier to conduct

Statistical significance

In addition to looking at the magnitude of the validity coefficient, the statistical significance should also be established. Statistical significance indicates the likelihood that your findings have occurred by chance rather than being a reflection of the true association between the variables.

During data collection, random variations in sampling mean that at times associations may be observed in the data when there is actually no association. This is particularly likely when small samples are used, as there is a much greater scope for error.

Statistical significance tells us the likelihood that a finding has occurred by chance, or alternatively, can be interpreted as indicating the confidence we can place in our results. To be meaningful, validity coefficients should therefore normally be at least 0.2-0.3 and also be statistically significant.

Using a significance table, you are able to look up the correlation coefficient for the number of participants which the coefficient of your study must meet or exceed in order to be considered statistically significant. Firstly, find the row that corresponds to the number of people used to calculate the correlation and then decide on the required level of significance, either 5% or 1%. Using a 5% level of significance means that we can be 95% certain that any finding represents a true association between the variables used in the validity study, whereas using a 1% significance level means we can be 99% certain of the results.

Example:

20 people in the validity study

Therefore, the correlation coefficient needs to be 0.444 or greater to be 95% certain the results reflect a true association

To be 99% certain the correlation coefficient must be 0.561 or greater

By looking at the following table you are able to see that for validity studies with very low numbers of people, high correlation coefficients need to be observed in order to be considered significant. As the numbers increase, there is a considerable drop in the coefficient needed as there is less scope for random error. Ideally there would be a minimum of 50 people included in a validity study for the findings to be robust and for a good chance of finding evidence of validity, if this exists.

Statistical Significance table

	Level of significance	
Number of people	**5%**	**1%**
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708
13	0.553	0.684
14	0.532	0.661
15	0.514	0.641
16	0.497	0.623
17	0.482	0.606
18	0.468	0.590
19	0.456	0.575
20	0.444	0.561
21	0.433	0.549
22	0.423	0.537

	Level of significance	
Number of people	**5%**	**1%**
23	0.413	0.526
24	0.404	0.515
25	0.396	0.505
26	0.388	0.496
27	0.381	0.487
28	0.374	0.479
29	0.367	0.471
30	0.361	0.463
35	0.334	0.430
40	0.312	0.403
45	0.294	0.380
50	0.279	0.361
60	0.254	0.330
70	0.235	0.306
80	0.220	0.286
90	0.207	0.270
100	0.197	0.256
200	0.139	0.182
400	0.098	0.129

Validity generalisation

The validity study results obtained from one situation cannot always be generalised or applied to others, meaning that test users cannot rely on existing research to guide their use of tests. Work by Schmidt and Hunter (1977) has suggested, however, that it is not always necessary to conduct validity studies for each application of a test as generalisation from validity studies can often be made.

Building on this initial work, later research has established that cognitive ability tests have validity across a wide range of jobs and organisations. In addition, results consistently show that predictive validity for cognitive ability tests is higher for more cognitively complex jobs. When jobs are grouped according to their complexity, predictive validity estimates for cognitive ability tests tend to range from approximately 0.25 for the least complex jobs up to around 0.55 for the most complex.

Whilst these findings do not contradict the importance of test users conducting their own validity studies, they do mean that existing validity research can be used to guide the selection of ability tests in new situations. When looking to existing studies to guide our selection of tests, the greater the similarity between the context of the study and our situation, the more confidence we can have that the results can be generalised. As a guide, the following criteria affect how confident we can be in generalising from the results of other studies:

- Similarity in job roles
- Similarity in performance criteria
- Similarity of tests used
- Similarity of applicants

Section 7 – Utility

Introduction

When a test is used, it should be because the information provided from that test will positively impact the individual and the organisation. Utility theory builds on validity by identifying these benefits. Within this section we will look at how the benefits from using tests or other assessment processes can be quantified.

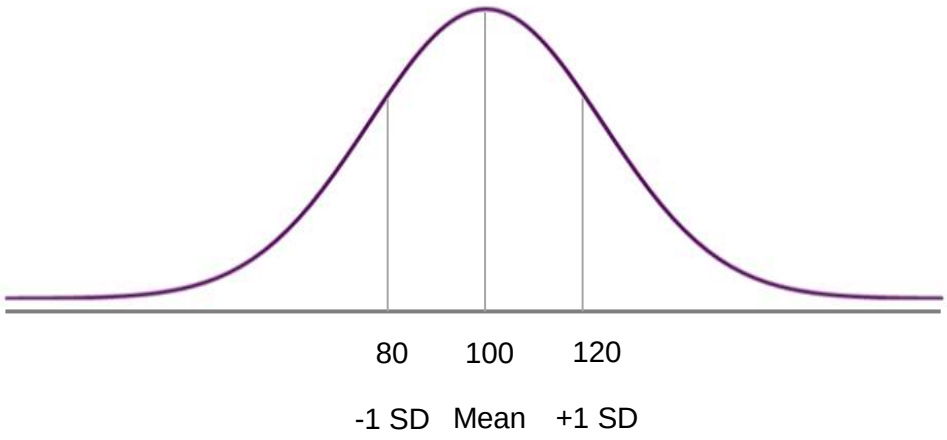
Predicting job performance

It is possible to use validity information and test scores to predict performance in the role.

Example:

First Class Office Supplies have been looking to employ a Junior Account Manager. They provide office supplies to a variety of businesses. Their existing Junior Account Managers' productivity is measured in number of sales.

The productivity is normally distributed across staff.



Productivity has been measured in units of £1,000. The mean ($\bar{Y}$) = £100,000 and SD_y = £ 20,000.

Candidates were tested as part of the selection process. Based on their performance on a Verbal Reasoning test we can predict their performance.

There are 28 questions on the test with a mean score ($\bar{X}$) of 20 and SD_x = 4 and the test's validity = 0.4.

Here are just five of the candidates who took part in selection.

Candidate	**Score (X)**
Brian	26
Laura	18
Caroline	23
Lynne	10
Harry	20

Once you know this information it is possible to calculate the prediction of their performance.

Firstly, we must calculate the z-score.

$Z = (X - \bar{X}) / SD$

Candidate	**Score**	**Z-Score**
Brian	26	1.5
Laura	18	-0.5
Caroline	23	0.75
Lynne	10	-2.5
Harry	20	0

To calculate the prediction of candidates' performance, we then multiply each individual's Z-score by the validity.

If we use Brian as the example,

$\hat{Z}_y = r \cdot Z_x$

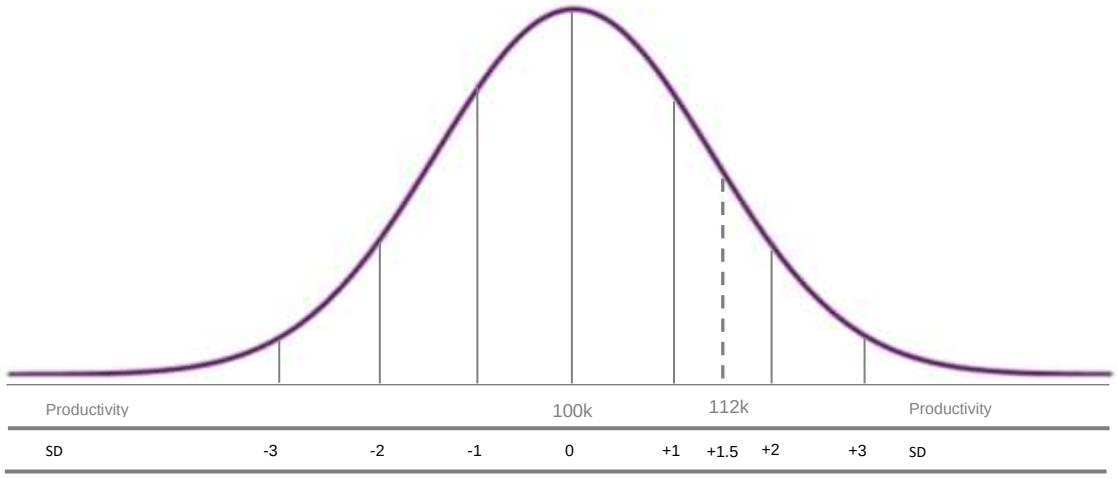
$= 1.5 \times 0.4 = 0.6$

This can then be converted to actual performance measure. To do this we multiply the Z-score prediction of performance by the standard deviation of the existing employees' performance. To this figure we then add the mean of the existing employees' performance.

$\hat{Y} = (\hat{Z}_y . SD_y) + \bar{Y}$

$= (0.6 \times £20{,}000) + £100{,}000 = £112{,}000$

So in Brian's case we would expect him to be more productive than the average employee and should generate **£112,000** worth of sales.

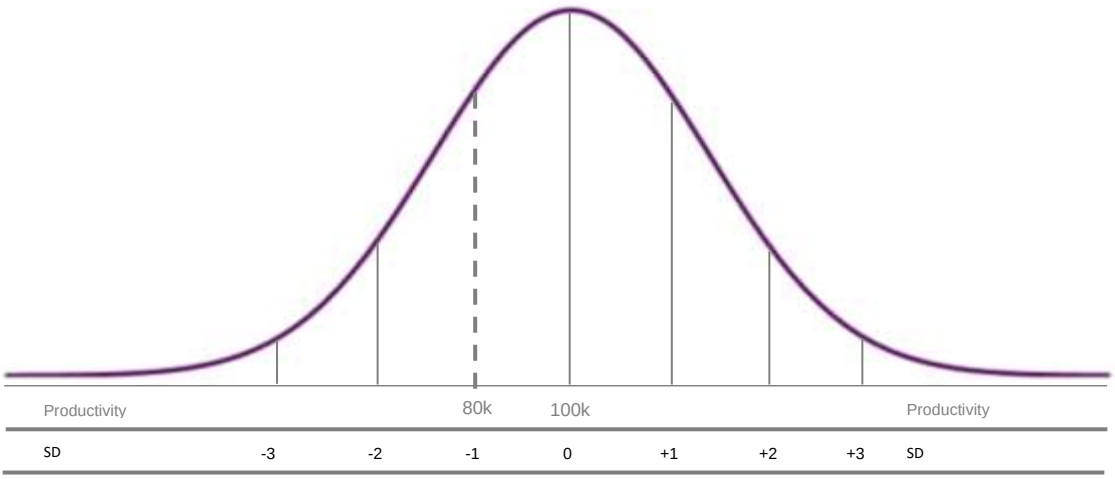


With a result like this we can be confident that Brian would be a good fit for the role. Whereas if we look at Lynne's test performance and convert that into a prediction of actual job performance we can see the difference.

$\hat{Z}_y = Z_x \times r \qquad = -2.5 \times 0.4 = -1$

$\hat{Y} = (\hat{Z}_y \times SDy) + \bar{Y} \qquad = (-1 \times £20{,}000) + £100{,}000 = £80{,}000$

As Lynne is only expected to generate £80,000 worth of income per annum you can see that she would be a less appealing candidate to employ.



Utility analysis

In order to calculate the financial gain of using psychometric tests Utility Analysis is used.

Cronbach's Utility Formula

£ gain = (r . $\bar{Z}_x$. SD_y. N) – C

Where N = Number of selected applicants

r = Validity coefficient

$\bar{Z}_x$ = Mean Z score on the selection test(s) of the selected candidates

SD_y = Standard deviation of profit on the job performance criterion in £

C = Cost of the testing programme

Example using the utility equation

We can also calculate the financial gain of using the tests in order to predict performance and aid selection using the formula in the box above.

We can use Brian from the example above to illustrate this point. If we multiply the validity coefficient of the test by the candidate's Z score (if you want to look at all the candidates together, you would use the mean Z score of all candidates) then you would multiply this by the SD of the existing employees' performance. If you were looking at a group of candidates score you would then multiply this figure by the number of candidates. The final stage of this equation is to subtract the cost of the testing process. This calculation identifies the financial gain of using the test to select candidates.

From our earlier example,

N = 1

r = 0.4

$\bar{Z}_x$ = 1.5

SD_y = £ 20,000

For the sake of this example we will say that the cost of testing the one individual, C = £50.

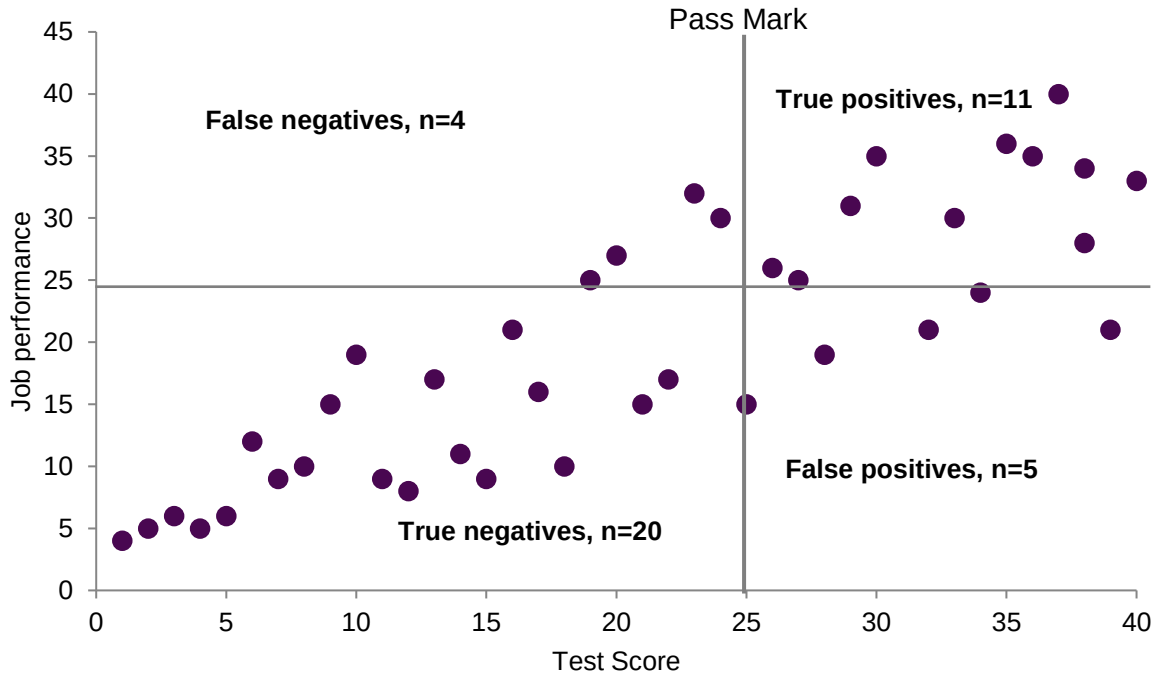
£ gain = (r . $\bar{Z}_x$. SD_y. N) – C = (0.4 x 1.5 x 20,000 x 1) – 50 = **£11,950**

Overall we can conclude that the financial gain of using testing in the selection of Brian is £11,950.

Alternative methods of interpreting validity evidence

Utility analysis provides a way of quantifying the benefits from using tests or other selection methods, but evidence of the link between test scores and job performance can also be used in other ways to support a selection process. An example of this is helping to identify where 'pass marks' should be set.

The scatter-plot below shows the association between test score and job performance. The dotted line indicates the desired level of job performance. This information can help identify where it would be most appropriate to set a 'pass mark'. When you set a pass mark you are intending to get as many people as possible in the top right and bottom left quadrants. Those in the top right are known as 'true positives', meaning that they meet the pass mark on the test and go on to perform at the required level. Those in the bottom left are known as 'true negatives' and fall below the pass mark and would not have met the required performance standard had they been selected. Effectively these are the 'correct' decisions. However, some people fall into the top left and become 'false negative', people who didn't meet the pass mark but would have performed at the required level, or they fall into the bottom right and are classed as 'false positives', people who met the pass mark but would not have gone on to perform at the required level.



Some incorrect decisions will always exist as no selection system can be perfect and free from error, but the goal will typically be to maximise correct decisions and minimise incorrect ones. As can be seen in the above example, setting a pass mark at 25 appears to be optimum, as most people are correctly identified as true positives or true negatives.

This formula can be used to calculate the actual accuracy of the selection process:

> Accuracy = (true positives + true negatives) ÷ total number of cases

In this example, Accuracy = (11+20) ÷ 40
Accuracy = 31 / 40 = 0.78 or 78%

Setting the pass mark at 25 therefore gives an accuracy or 'hit rate' or 0.78 or 78%. Using the data from the scatter-plot, the accuracy can be calculated for different pass marks until the optimum pass mark is identified.

By using multiple methods as part of the selection process this can compensate for any gaps in the accuracy. It can help improve the accuracy of the final decision by triangulating the test data with that of the additional information gathered from other methods.

Section 8 – Test Administration

Introduction

This section will cover all the areas of test administration including the stages of preparation and the arrangements to be made prior to testing as well as the key principles of test administration itself. This section will also cover the structure of the test session and post-session arrangements.

In order for tests to be an accurate assessment of an individual's ability or other characteristic it is essential that they are administered in accordance with standardised procedures and instructions. Good test administration is a result of careful preparation and a good understanding of the test materials. It is also aided by the administrator developing a rapport with the test-takers and providing appropriate support prior to, and during, the test session.

Test administration process

Testing is typically used as part of selection or a development activity, meaning that it will not usually occur in isolation but will be part of a broader process.

Administering a test or battery of tests requires careful planning, and can be helped by breaking down the test session into four stages:

Once you know which test is going to be administered you are able to plan the testing session around it and any other assessment activities that the test-takers may be required to undertake. Planning a test session is very similar to any other form of event planning. The checklist on the next page details the points you need to be able to answer and the areas which need to be considered in order to be fully prepared.

Preparation checklist

Preparation Checklist	
Date of test session	
Which tests are being used?	
How long will be needed for the test session?	
What test materials/equipment is required?	
Will the administrator need any assistance (e.g. due to large numbers of test-takers)?	
When and how will the test-takers be informed about the test session?	
Will the test-takers receive any preparatory materials?	
Will the test-takers be offered feedback?	
How and in what format will the test scores be communicated to decision-makers?	

Arrangements prior to testing

Timing of the test session – In advance of conducting a test session, it is important to be clear about which tests are to be taken and how long these will take to complete. There are other factors than just the test time limit when determining the length of the test session. You also need to take into account the time required for the informal introduction, reading the test instructions and attempting the practice questions, and for the distribution and collection of test materials.

Arranging a suitable venue – Once you know how long you need for the testing session you will be in a position to be able to find a suitable venue. The venue must be large enough to comfortably accommodate all test-takers and be free from distractions which may affect the test-takers' concentration. In addition to the actual room, sufficient desks and chairs should be available along with any other equipment you may need.

Inviting test-takers to the session – When you know the arrangements for the day, you need to contact test-takers and invite them to attend the test session. There are several points which need to be included when contacting test-takers whether it is by letter or email or for a testing session or part of a broader assessment day. These points are listed in the checklist below.

Invitation Checklist	
Purpose of the testing session	
Which tests will be completed	
Venue and timings	
Anything the test-takers should bring with them (e.g. reading glasses)	
How they can best prepare for the tests	
How their results will be used	
Confidentiality and who will have access to the results	
How they will receive feedback	
Who to contact if they have any special requirements or queries about the day	

Example test session invitation letter

Dear Marie

Further to your application for the position of Junior Account Manager at Class One Office Supplies, I would like to invite you to attend an assessment day starting at 9:30 a.m. on 15 May. The day will be held at our offices at the address given above. The purpose of the assessment day is to help us to assess your suitability for the role of Junior Account Manager. The day will be split into two halves, with psychometric testing in the morning followed by an interview in the afternoon.

During the morning you will complete the following three timed psychometric ability tests:

Saville Consulting Commercial Verbal Comprehension – a test which assesses your ability to check the accuracy of written material.

Saville Consulting Commercial Numerical Comprehension – a test of your ability to check the accuracy of numerical material and correct any mistakes.

Saville Consulting Commercial Error Comprehension – a test of your ability to check the correctness of transposed information.

Full instructions for each of these tests will be given on the day and I enclose a leaflet explaining further about the tests and how you can best prepare for them. Please note that the tests will require attention to written material, so please bring reading glasses if you require them.

The results from these tests will be used in conjunction with your interview as part of our selection procedure. Feedback on the test results will be available.

I would like to reassure you that the results will remain strictly confidential and will be seen only by myself and the members of the interview panel. All test results will be stored in accordance with the Data Protection Act. At Class One Office Supplies, it is our policy to store all test results for six months, after which they will be anonymised and used for monitoring and research purposes only.

If you have any special requirements, I would be grateful if you could contact me in advance of the assessment day.

Yours sincerely,

Making adjustments to psychometric testing – This can be a complex process as changes can put the standardisation of procedures at risk. Changes should only be made after consulting both the individual test-taker concerned as they will be the best source of information on what adjustments are appropriate for them and the test publisher. Adjustments can take time to put in place so it is important that the test users have as much notice as possible; therefore, test-takers need to be asked if they have any special requirements well ahead of the test session itself.

Test materials and other equipment – In advance of the test session, checks should be made to ensure that all the necessary test materials are available. What materials are needed will depend on how tests are being administered and whether they allow the use of equipment such as rough paper or calculators, but administrators should make sure that there are sufficient materials for the anticipated number of test-takers plus a few spares.
Reusable question booklets should be carefully checked if they are being used as previous test-takers may have made marks in them. If marks cannot be completely erased replacement books should be ordered. If the tests are to be delivered by computer, each device should be checked before the test session to see that they are working appropriately and any necessary software is installed and working.

Managing the test session – When planning a test session, administrators should consider how many people they will be testing at one time and so whether they may require an assistant to help them with distribution and collection of test materials. It is particularly important that test-takers do not start looking through the booklets when they are being administered until they are told to do so. As a guideline, one administrator can work with groups of around 10 to 12, but with larger groups a second person is required.

Preparing for the test session

Shortly before the testing day it is advisable to recheck the venue and any other arrangements you have previously made including the stock of testing materials.

For the day of testing, the room should be arranged so that test-takers are far enough apart so that they cannot see each other's paper or inadvertently distract each other. Test-takers should have sufficient space on their desks for any materials they will need and there should be space between the test-takers' desks so that the administrator can walk comfortably between them. The room should be well lit, though be free of glaring sunlight, and should be at an appropriate temperature.

Using test logs – Test logs are a convenient way of helping a test administrator prepare for a test session and providing a written record of the session.
Test logs should be completed when preparing for the test session and as part of the session itself, ensuring they provide an accurate record of the process. Being a record of actual events, they act as

an audit to ensure that testing was conducted appropriately and all candidates had an appropriate opportunity to demonstrate their abilities during the session.

Informal introduction – The informal introduction gives the test administrator the opportunity to provide test-takers with information about the test session in their own words and put them in an appropriate mind-set for completing the test. It should be delivered in a friendly and efficient manner, reduce any anxieties test-takers may have and result in them being motivated and ready to start the test.

Administrators should be aware that in recruitment situations this may be one of the first face-to-face contacts that test-takers have with a potential employer. In addition to conducting the test session effectively, administrators should be aware that how the session is conducted is likely to impact on the test-takers' views of the organisation.

Administrators should prepare their informal introduction in advance of the test session and check that it covers all necessary points. You should prepare in whatever way you feel comfortable; whether this is writing it out in full or using crib sheets. An informal introduction should include the following points:

Introduction Checklist	
Name of the administrator and their role in the organisation	
The purpose of the test session	
Prompt test-takers to turn off mobile phones and other distractions	
How the tests relate to the role or other activities	
Which tests will be taken and what they assess	
How the results will be used	
Who sees the results	
How the results will be stored	
Opportunities for feedback	
That the tests are strictly timed and therefore must work quickly and accurately	
That fixed instructions will be given to ensure standardisation	
The opportunity for test-takers to ask any questions	

After any questions have been answered, it is advisable to get the test-takers' consent to proceed by asking something to the effect of 'Is everyone happy to proceed to the formal part of the test session?'

Formal instructions – To help a test session run smoothly the administrator needs to be familiar with the formal instructions. They need to know what instructions need to be read out to the test-takers, when answer sheets and question booklets should be distributed, and any other aspects of the testing procedure. It is also important for administrators to have a clear understanding of the solutions to any practice questions, so that they can provide any assistance that is permitted if test-takers have queries about these.

Standardisation has been identified as one of the key properties of psychometric tests, and administration is one of the major areas through which standardisation is achieved. All psychometric tests will come with guidance on how they should be administered to ensure consistency and standardisation. An important part of this guidance is a set of formal administration instructions which will usually contain:

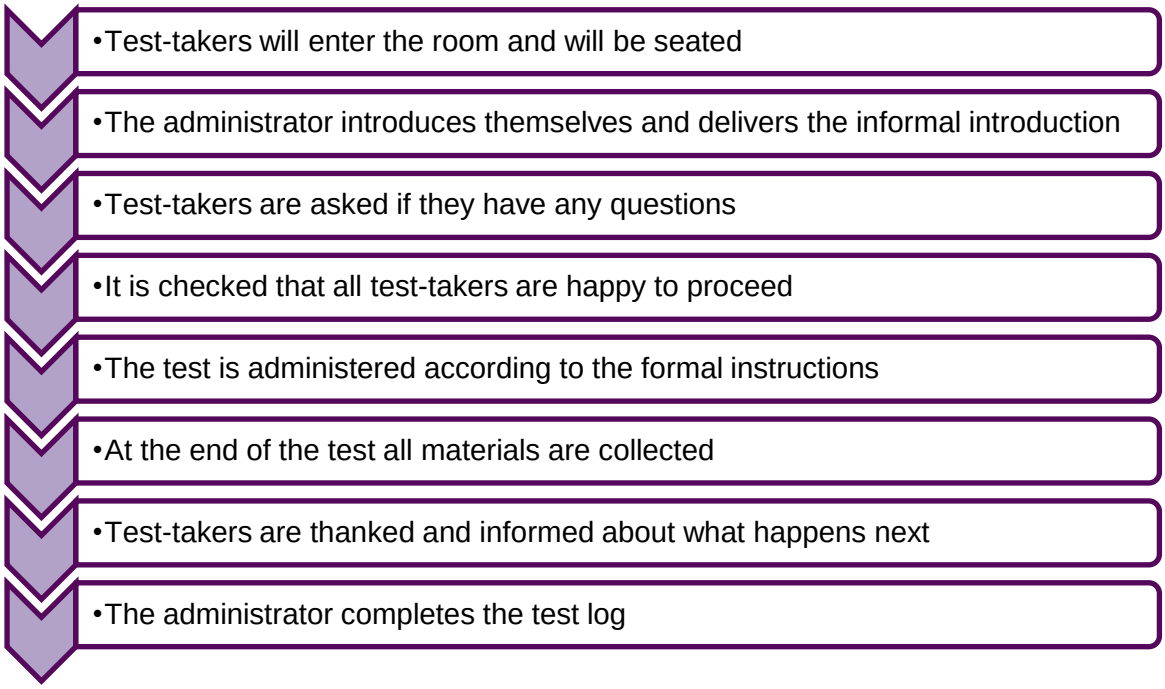
- Test instructions that the administrator reads to test-takers
- Guidance to the administrator on the process (e.g. when to hand out answer sheets and question booklets)
- Other tips such as handling the practice items and any questions that delegates have
- The exact timing for the test

The best way to understand the requirements of a test is to take it yourself. Though most modern ability and aptitude tests follow a similar structure, no two tests are identical in their administration. Believing that you can administer a new test because you have used similar ones in the past is very dangerous; you do not want to be caught out.

Final checks – On the day of testing you should check the room where the testing will be conducted and any equipment such as computers which will be used. All materials including name cards, rough paper and pens should be set out as appropriate prior to the candidates entering the room.

The test session itself

A testing session usually follows a simple structure:



Dealing with questions

Test-takers are given the opportunity to ask questions during test administration. As an administrator you need to be prepared to answer any questions, though responses should be succinct as the administration session is not the place to enter into extended discussions with the test-takers.

Common questions asked by test-takers:

Why are you using these tests?

Who will see the results?

Do I have to do these tests?

Is there a pass mark?

What will happen to the results if I am successful / unsuccessful?

Should I skip questions if I get stuck on them?

Are these tests fair?

I have not taken tests like this before, so will I be disadvantaged against others who have?

Is it better to work slowly and accurately or should I focus more on getting through the test?

Many of these questions can be addressed through a good informal introduction but despite this at times there will be some questions which are not covered by the introduction. When answering questions, it is important that administrators are able to give concise answers or, if this is not possible at the time, make a note of the question and provide an answer to the test-taker after the session.

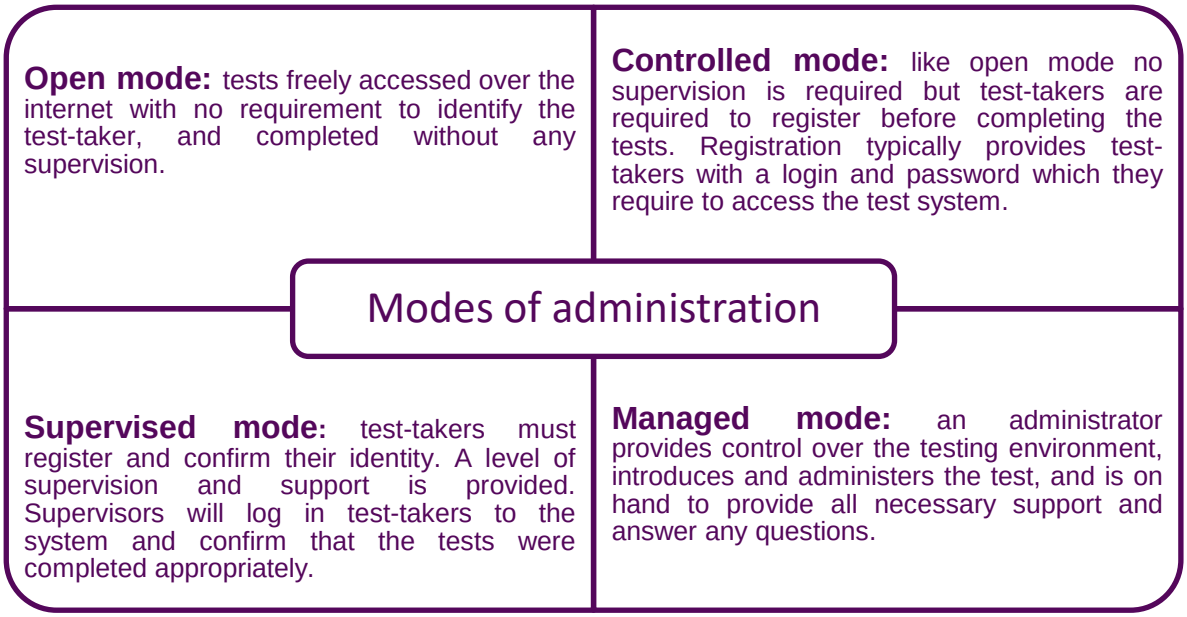
Computer-based testing

The test administration process described so far has assumed that paper-and-pencil materials are being used. However, nowadays computer-based testing, usually over the Internet, is more popular. Computers can be used to administer tests, score them and generate reports of the results. Computers bring significant advantages to the testing process, including complete standardisation in the delivery of the instructions, ergonomic advantages (e.g. no need to work with separate question booklets and answer sheets), precise timing, accurate scoring and consistent interpretation.

When computers are used to administer tests under supervised conditions, this is unlikely to have a significant effect on the administration process. It is the increasing use of the Internet that brings additional consideration into the administration process. The main reason for this is that the Internet allows tests to be administered remotely under unsupervised conditions. When planning a test session using internet-based testing, the appropriate 'mode' of administration needs to be considered.

Four modes of administration

Bartram (2000) defined four modes of test administration which have been incorporated into the International Test Commission's (ITC) Guidelines on Computer-Based and Internet Delivered Testing (2005):



	Advantages	Disadvantages
Open	• Low costs • Easily accessible • Good for informal purposes • Remote access	• Cannot authenticate candidate identity • Non-standardised environment • Lack of security of test items
Controlled	• Test-takers identity is known • Control over access to the test • Remote access • Relatively cheap • Can be used for formal testing	• Potential for cheating • Non-standardised environment • Requires subsequent re-verification of performance
Supervised	• Greater control of candidates • Can authenticate candidate identity • More standardised environment • Can be used for formal testing	• Requires someone to be present • Cannot be done remotely • Relatively more expensive
Managed	• Highly standardised environment • More control over the equipment being used • High level of control over candidates • Good for formal testing	• Requires someone to be present • Requires a dedicated testing centre • Can be expensive

Checking candidate behaviour

As mentioned above, there is the potential for cheating when administering tests using the controlled mode of administration. If a candidate completes a test in a remote setting, there is a chance that they have had assistance or cheated when completing the test. In order to overcome this obstacle, organisations can use a variety of methods to check candidate behaviour remotely.

Follow-up testing

Some organisations used the controlled mode of administration for high volumes of candidates in the early stages of the selection process. This allows them to short list candidates based on their performance on the initial tests. Typically, this is then followed by a subsequent testing session under supervised test conditions for a much smaller group of candidates to verify the performance of the candidate.

Forensic data analysis

Forensic data analysis is used to assess the trends in data in order to assess whether the behaviour displayed by a test-taker is atypical and therefore implies that they may have attempted to manipulate the results or have behaved in an unethical manner. For example, if someone chose to write down each of the questions there would be a long time lag for each question. The analysis would flag this as something which may require further investigation.

Section 9 – Feedback

Introduction

For whatever reason a test-taker has completed a test, best practice dictates that feedback on their results should be made available to them. Such feedback is common in development contexts as the reason for using the test will have been to increase self-awareness in the test-taker to help them with their personal development. In selection contexts, practice in giving feedback is more variable, with some organisations ensuring that everyone can access feedback whilst others do not offer it.

Reasons for giving feedback	**Possible reasons why feedback is not offered**
Recognising that testing should be a two-way exchange of information	The time and cost involved, especially when tests are being used to screen large numbers of people
Enabling test-takers the opportunity to use the test results to learn and develop	Lack of suitably qualified people to give feedback
To enable a discussion around the results that may give the organisation further valuable information on the test-taker	The belief that knowledge of results is not particularly useful to test-takers
Creating a positive impression of the organisation	The belief that test-takers can't change their abilities, and so feedback is of little use to them

Access to information under the Data Protection Act

When an organisation holds test data in a way that can be linked back to individual respondents, the individual can gain access to their data under the Data Protection Act. If test-takers are offered feedback, they are given the opportunity to discuss the test data. This is the most common way of seeing data held whereas requests to access the data is far less common. A feedback situation is beneficial for both parties and a request for information would be time consuming and disruptive, these are just some of the reasons why it is good to offer feedback to test-takers.

Gathering the necessary information for feedback

Before the results can be communicated to any of the stakeholders involved in the testing process, the test users need to make sure that they have all of the necessary information available.

This information is likely to include:

- A clear understanding of the purpose of testing
- Which tests have been used and what each of them measures in non-technical language
- The test-taker's raw score on each test, the number of questions in the test, number attempted and number correct
- The norm group that has been used to interpret the raw score
- The test-taker's standardised scores taken from the norm group

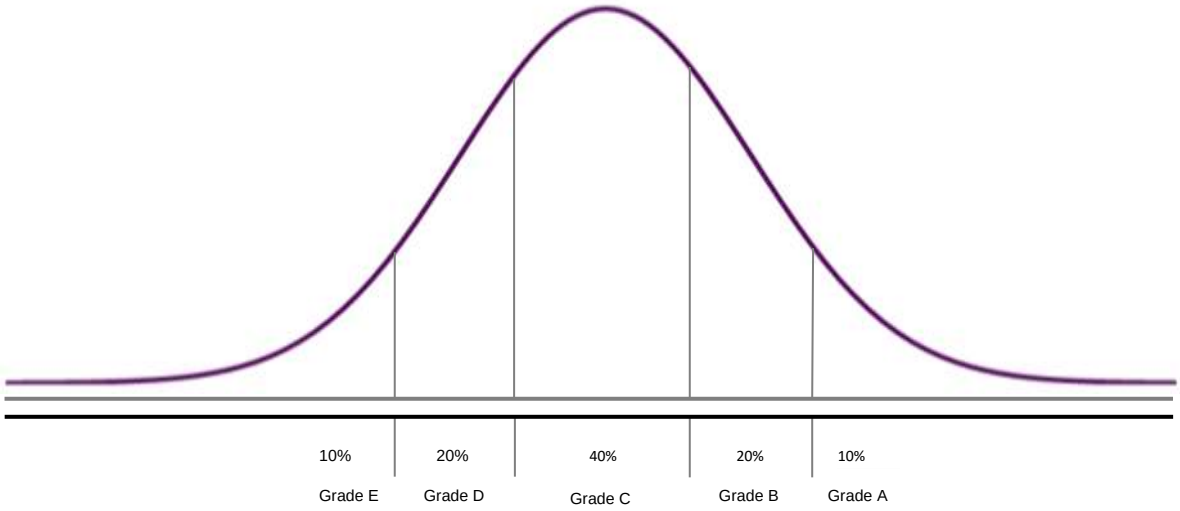
Once this information is gathered it can then be communicated in an appropriate format to the stakeholders involved in the testing process. Most of it is necessary to accurately communicate the findings to all parties even though some of the information will not be communicated in its 'raw' form. For example, the SEd would not be communicated directly to test-takers, but knowledge of this is important if we are to say whether a test-taker's results on one test are significantly different from their results on another test.

Communicating scores to test-takers

You need to consider the format which will be used to express the results when communicating the results to someone who is not trained in psychometric testing.

Feedback is used to ensure that the test-takers have a clear understanding of their performance and its implications. There is no set format for expressing test scores but there is a general consensus that standardised scores such as Z scores and T scores are open to potential misinterpretation and so should not be used in this context. Instead, users may choose to use 'bands' to communicate scores or use percentiles if the actual scores are being communicated.

Score bands may be expressed in a number of ways, though all attach a description of the score depending on where they fall in relation to the norm group.



In this example, grades have been attached to the bands of percentiles. The letters A to E have been used to indicate the standard performance corresponding to five splits of the score distribution.

Verbal descriptions of the scale bands may also be given. The table below illustrates the kind of verbal description which would be attached to each of the grades. Other terms may be used, particularly for organisations that use competency frameworks, these tend to use language based in competence or development needs. For example, percentiles between 81 and 99 may be described as 'highly competent' or an area of considerable strength', whereas percentiles of 1-10 may be termed 'an important development need'.

Percentile band	**Grade**	**Description**
1-9	E	Well below average
10-30	D	Below average
31-70	C	Average
71-90	B	Above average
91-99	A	Well above average

As with any score based on a normative interpretation, it should be clearly communicated to test-takers that the grades or verbal descriptions of scores are given relative to a specific norm or comparison group.

Percentiles – If users want to communicate more precise scores, the most appropriate format for doing this is using percentiles. Percentiles have the advantage of being more easily understood and are simpler to communicate than standardised scores.

In addition to the actual percentile score, test-takers also need to appreciate the following points:

- That the score is a ‘percentile’, NOT the ‘percentage correct’
- As with all scores, it is interpreted relative to the norm group and so is not fixed
- That due to error in measurement, the percentile is an indicator of their ability rather than an absolute value, and so should not be over-interpreted.

Qualitative interpretations of results

As well as using score bands or percentiles sometimes it can be helpful to the test-taker to have a more qualitative interpretation of their results. A qualitative analysis can help the test-taker to explore the approach they took to completing the test by considering the speed and accuracy of their working. In order to do this those providing feedback need to know the total number of questions in the test, how many the test-taker attempted and how many of these were answered correctly.

Preparing for feedback

Good preparation is an important aspect of giving successful feedback. A time needs to be set in advance whether the feedback is going to be conducted face-to-face or over the telephone. For face-to-face feedback, a suitable room, free from disruptions, also needs to be arranged. Resources such as paper and pencils may also need to be provided therefore these must be made available in case the test-taker wants to make notes about their results.

Most of the computer-based testing systems will provide you with the option to produce a report suitable for giving feedback directly to test-takers. If these, or reports produced manually, are to be given to test-takers as part of the feedback session, these also need to be available. A decision needs to be made as to when such reports should be given to test-takers. It is recommended that feedback reports are given either before the feedback session or at the end of it, as if reports are given during the session, test-takers may be distracted by the reports and not fully engage with the feedback conversation.

If the feedback is part of a development process it can be useful to give the test-taker the report prior to the feedback in order to provide them with the opportunity to come to the feedback session with any questions or points they wish to discuss during the session. For selection it might be better to give the reports at the end of the session particularly if there is limited time for the feedback.

Conducting a feedback session

The structure of the feedback session is dependent on the purpose for which the tests have been taken. A session used to explore the implications of tests taken as part of a development would be quite different from a session for feeding back the results from tests used as part of a selection process.

The model provided here outlines the basic structure of a feedback session in four parts:

This is a widely used model and offers a robust and logical framework for communicating the results from any test or assessment, though what goes in to each of these sections will depend on the purpose for which the assessments are being used.

Introduction

In this section you need to outline the purpose of the feedback session, allowing you to manage the expectations of both the test-taker and the person delivering feedback around the contents and outcomes of the feedback discussion.

The main points that should be covered in the introduction are:

Welcome	Welcome the test-taker to the feedback session and an introduction by the person giving feedback
Purpose	Explain the purpose of the session is to review the results from the tests and, where appropriate, state what outcomes are expected from the session
Context	A brief reminder of why the tests were taken and their role in the selection or development process
Two-way discussion	Emphasise to the test-taker that the session should be a shared discussion, not just a delivery of the test results
Confidentiality	Explain the confidentiality of the feedback session and what will happen to test results after the session
Time limits	Give the test-taker an indication of approximately how long the feedback session will last

Communicating test results and exploration

This part makes up the main body of the session and is where the results of the test are communicated to the test-taker. The implications of the test results could be explored; the depth of this exploration is dependent on the purpose of testing. If the feedback is part of a development process the depth of this exploration is likely to be much greater.

The table below shows the basic information which would be communicated in all contexts with the level of exploration adapted according to the context:

Explanation of the norm-referencing process	In simple, non-technical language, describe how the test results are put into context
Description of the results on each test	Clearly describe the test results in relation to the norm group (e.g. as a score band or a percentile), ideally express the results in relation to the test-taker's expectations of their performance
Exploration of the implications of the test results	Question the test-taker to understand how these results relate to their experiences (e.g. current job role, education, interests) and development needs and career opportunities
Review of qualitative interpretation of results	Where necessary an exploration of the speed and accuracy of working to help explain any unexpected results and potential strategies for development

It is important to be sensitive throughout the feedback to the reactions of the test-taker, therefore when giving feedback you need to be aware of both verbal and non-verbal cues. If the feedback is being given over the telephone, the non-verbal cues will not be visible so it is vital that you are particularly aware of the verbal cues to the emotional reactions to the results.

Summary and close

The final section of the feedback involves bringing the information which has been discussed together into a succinct summary.

Written feedback

If the feedback is going to be communicated in a written format the way in which the results are expressed needs to be carefully considered. The information needs to be written in a manner which can be easily understood as there will not be a trained test user present to provide additional clarification. Many of the guidelines for verbal feedback can be applied to written feedback.

There is an example of a written feedback report in the workbook.

Appendix

A1 Examples of tests of reasoning ability

Once you have the Test User: Occupational, Ability (Level A) qualification, you can use any of these tests. You simply need to contact the test publisher who will ask for confirmation of your qualification. The table below contains just some of the tests which are available.

Publisher	Test name
CEB - SHL www.ceb.shl.com	Management & Graduate Item Bank (MGIB) – consisting of Verbal and Numerical Critical Reasoning Advanced Managerial Tests (AMT) – consisting of numerical and verbal reasoning Verify – consisting of Verbal, Numerical and Inductive Reasoning Graduate and Managerial Assessment (GMA) – consisting of Verbal, Numerical and Abstract Reasoning Critical Reasoning Tests (CRT) – Consisting of Verbal and Numerical Critical Reasoning Tests
Pearson Assessment www.talentlens.co.uk	Differential Aptitude Tests (DAT) – consisting of Verbal, Numerical, Abstract, Mechanical Reasoning and Space Relations Ravens Advanced Progressive Matrices (APM) – consisting of Abstract Reasoning Watson-Glaser Critical Thinking Appraisal – consisting of Verbal Reasoning
Saville Consulting www.savilleconsulting.com	Analysis aptitude, Comprehension aptitude, Technical aptitude Professional and Work aptitudes
Criterion Partnership www.criterionpartnership.co.uk	Utopia critical reasoning tests B2C Critical Reasoning Tests Criterion Workforce Series (CWS) critical reasoning tests

A2 Examples of personality tests

In order to interpret and provide feedback on these tests you will require the Test User: Occupational, Personality (Level B) qualification and usually specific training in each tool you wish to use. If you are not qualified, you can contact Psysoft and we can administer and interpret these tools for you.

Publisher	Test name
MHS www.mhs.com	Pearman Personality Integrator EQ-i 2.0, EQ 360 (emotional intelligence)
OPP www.opp.eu.com	Myers-Briggs Type Indicator (MBTI) 16 PF
CEB - SHL www.ceb.shl.com	Occupational Personality Questionnaire (OPQ) Motivation Questionnaire (MQ)
Saville Consulting www.savilleconsulting.com	Wave Personality Questionnaire
Hogrefe www.hogrefe.co.uk	NEO PI-R
Criterion Partnership www.criterionpartnership.co.uk	CAL Online
Hogan Assessments www.hoganassessments.com	Hogan Personality Inventory (HPI) Motives Values and Preferences Inventory (MVPI)
Insights www.insights.com	Insights Discovery