Defining the Assessment Needs

Introduction

This unit deals with the essential psychometric concepts of reliability and validity. After a brief overview, the statistical concept of correlation is introduced to help you understand how some measures of reliability are estimated and expressed. Detailed information is then given on how to measure the reliability and validity of a test, in order to better understand these concepts.

The unit will:

- ☐ introduce you to the concepts of reliability and validity, which give you an indication of the quality of tests;
- ☐ introduce you to the concept of correlation, which will help you understand how the reliability and validity of a test are estimated;
- ☐ describe the different types of reliability and validity and explain how these are measured; and
- ☐ help you to assess tests and make good choices about test use.

Competencies covered in this unit

3.1 Explain the notion of correlation as a measure of the degree of relationship between two measures.

3.2 Define the conditions under which a correlation is maximised (both positively and negatively) and minimised.

3.3 Provide reasonable rough estimates of the correlation coefficients represented by examples of various bivariate scattergrams.

3.4 Explain in outline the methods of estimating reliability (inter-rater reliability, internal consistency, test-retest, same or alternate form) and describe their relative pros and cons.

3.5 Describe and illustrate the distinctions between face, content, construct domain and criterion-related validity.

Introduction to Reliability and Validity

Before we move onto the first competency of this unit, we will briefly introduce the concepts of reliability and validity. A good understanding of these concepts will enable you to:

- ➤ explain variations in scores between test scores;
- ➤ compare tests and choose wisely;
- ➤ make sound purchasing decisions;
- ➤ treat test scores cautiously and not over-rely on the exact test scores; and
- ➤ justify not using a test.

We'll start with a basic premise of Classical Test Theory (see also the Classical Model of Test Error in A Psychometrics Primer by Paul Kline, pages 42 to 45). This theory (or model) states that an observed score always has an inherent degree of error. In other words,

observed score = true score + measurement error

The aim of every designer of psychological tests is to minimise measurement error, so that the true score (or reality) can be accurately measured. The lower the measurement error, the more reliable and valid the test is.

Reputable test publishers provide information about reliability and validity to help inform test purchasing decisions. This unit will help you understand which information should be provided, when to ask more questions when certain information appears to be missing and notice when information is being presented in its best possible light!

A test that has a high degree of measurement error is not worth using as it will not be able to detect small differences in scores. For example, let's say my weighing scales at home are a bit 'dizzy' and any reading fluctuates by up to three pounds up and three pounds down (i.e. if my true weight is 10 stone 6 pounds, the scales will give me a reading that could be anything from 10 stone 3 pounds to 10 stone 9 pounds). I might lose two pounds (i.e. my true weight will then be 10 stone 4 pounds) but I might not be able to tell from my scale reading (i.e. the possible range of readings will overlap with the previous one), which will be really discouraging after having starved myself all week!

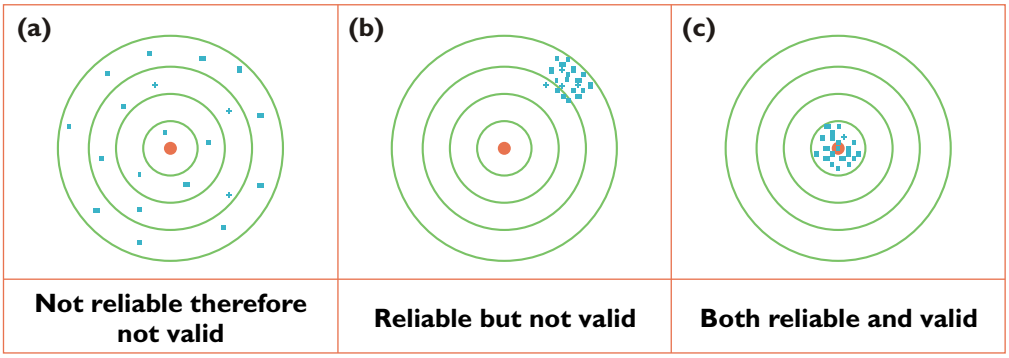
It is important to use reliable and valid tests and not only those which have been glossily and professionally produced. It is also important to note that just because a test has been standardised on a large population this does not mean a test is reliable and valid. The 'dizzy' scales above could be standardised on 2000 people but that would not make them less 'dizzy'!

So what's the difference between reliability and validity?

Reliability refers to how consistent or repeatable a test is whereas validity is about how confident we are that the test is measuring what it claims to measure.

Reliability is usually affected by what we call 'random error'. Random error can be due to many different aspects that are specific to the individual taking the test. These include fatigue and boredom. As can be deduced from its name, random error is assumed to affect people randomly, that is, some people might feel particularly tired on the day of testing while others might feel quite energetic that day. In this way, random error causes an increase in the variability of scores. An illustration of this effect is shown in Figure 3.1 (a), where the *true score* of the variable we want to measure is represented by the bull's eye and each dot represents a test *observed score*. The gap between the true score and the observed score is the *error*.

Figure 3.1 **The relationship between reliability and validity (adapted from Trochim, 2002).**



Validity, on the other hand, is mainly affected by 'systematic error'. This type of error is not associated with individuals, but with the environment or the test itself, and usually creates a consistent increase or decrease in scores. For example, there might be a lot of building work noise while a group of participants is completing a mathematical test. This noise could have a detrimental effect on their performance so that most people will score lower than they would have done in a quieter environment (which would have reflected their true score more accurately). Figure 3.1 (b) is an example of the effect of systematic error.

It is important to note how reliability and validity relate to one another. In Figure 3.1 (a), we had a test that wasn't reliable, because the scores obtained deviated too much from the true variable (or score) to be measured (i.e. represented by the bull's eye). Because they were so far apart (i.e. low reliability, high degree of random error), we couldn't really claim that our test was measuring what it was supposed to measure (i.e. bull's eye) and therefore lacked validity. In other words, if you have a test that is not reliable, it follows that it won't be valid either.

However, while reliability is a ***necessary*** condition for validity, it is not ***sufficient***. For example, in Figure 3.1 (b), we have a test that is highly reliable (i.e. measurements are consistent), but that is not really measuring what it intends to, since it's well off the bull's eye (i.e. high reliability, low degree of random error, but high degree of systematic error).

An example of a test that is both reasonably reliable (i.e. low in random error) and valid (i.e. low in systematic error) is presented in Figure 3.1 (c). Note that, just as in real life, there is still some test error.

For further definitions of Reliability and Validity type "define:reliability" or "define:validity", without the inverted commas, into the web site www.google.com. This will yield a range of excellent definitions.

Defining and measuring correlations

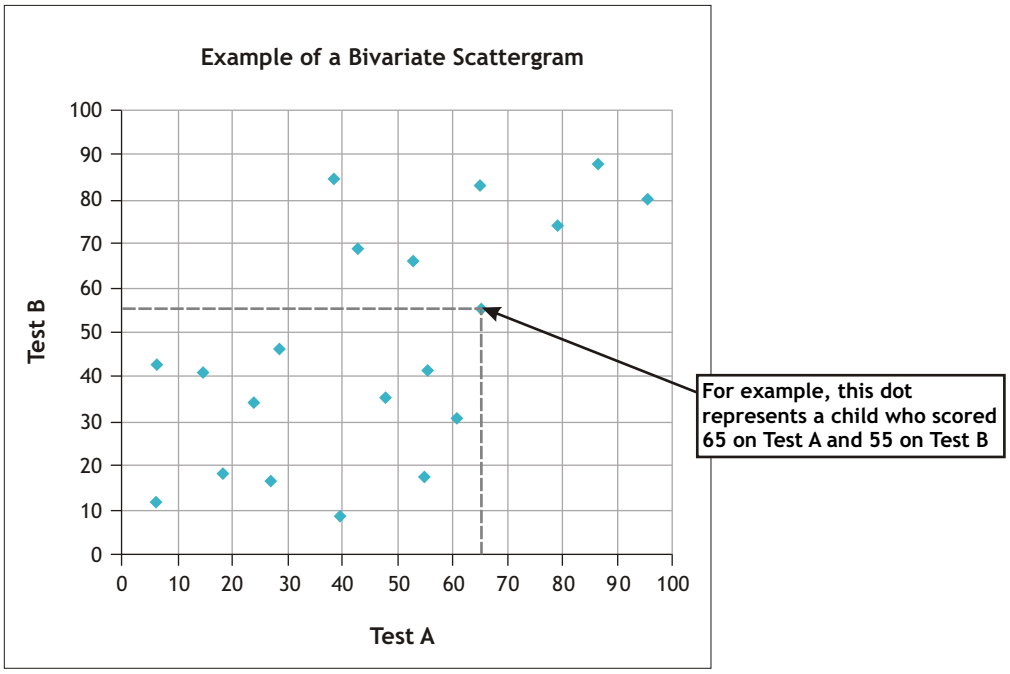
As we said earlier, you need to know about the statistical concept of correlation in order to understand how some forms of reliability and validity are estimated and expressed.

For example, when you look in test manuals, a figure will be given for reliability and in order to understand whether this is a figure you can trust it is useful to know how it was calculated.

A correlation is an expression of the degree of relationship between two variables (that's why it is called a '*bivariate* correlation').

In graphical form, this relationship usually takes the form of a scattergram. For example, let's say we have a group of 20 children and we get them to complete two separate tests that assess reading comprehension, which we will call Test A and Test B. If we were going to represent these scores in a scattergram, we will create something like the graph shown in Figure 3.2. Each dot represents a child and this is positioned in the graph according to the two scores obtained.

Figure 3.2 **Example of a bivariate scattergram showing the relationship between scores on Tests A and B.**



By looking at this scattergram, it is possible to make a judgement about the strength of the relationship between the two tests. They may claim they measure the same thing (i.e. reading comprehension). In this case, there appears to be a moderate relationship between the tests scores. If someone scored high on one test, they tended to score relatively high on the other test. However, that was not always true. If both tests claim to accurately measure exactly the same thing, then something is wrong with the reliability or validity of either or both of the tests. In a perfect world full of perfectly predictable children (just imagine!), perfect tests and perfect test administrators then a score on one test would correlate exactly with the score on the other test and the graph would be a straight line. See Figure 3.3(b) for an example of a perfect correlation.

TASK 24

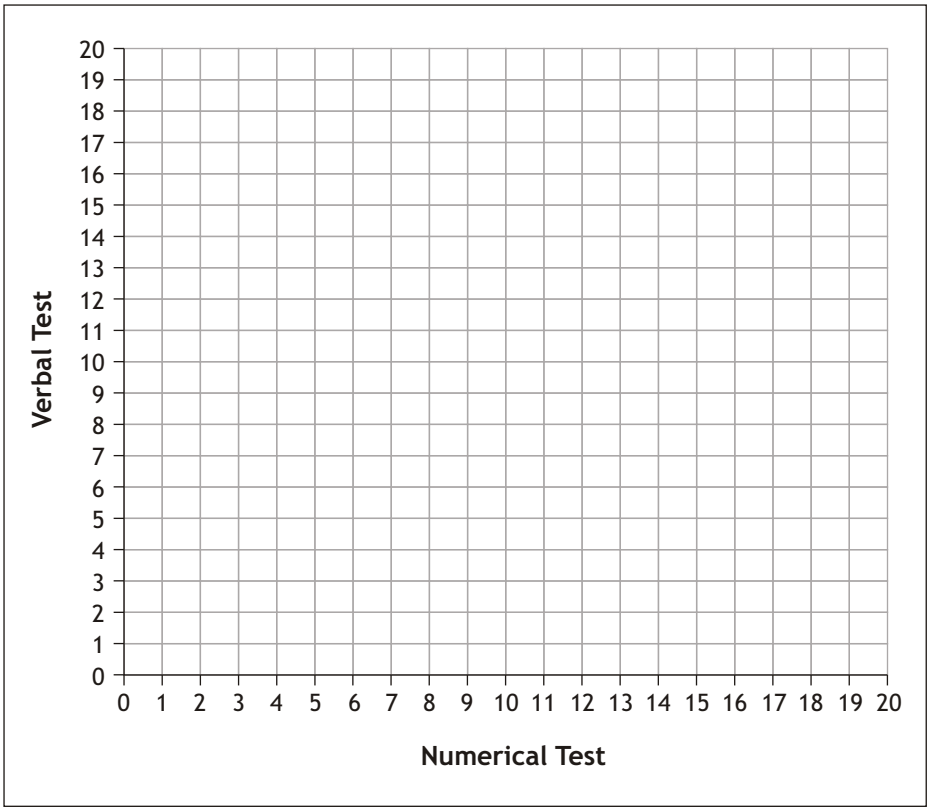
This task includes a couple of exercises that will help you to accurately interpret scattergrams and understand how these are created.

Answer the following questions based on the information given in Figure 3.2.

How many children scored between 20 and 30 in Test A?	
How many children scored between 20 and 40 in Test B?	
Did anyone score between 10 and 20 in both tests?	
Did anyone score between 30 and 50 in both tests?	
The highest scorer for Test A, what score did s/he get for Test B (approx.)?	
The lowest scorer for Test B, what score did s/he get for Test A (approx.)?	

Twenty children complete both a numerical and a verbal task. Using the scores given below, create a scattergram to represent these scores.

	Numerical Test Score	Verbal Test Score
Child 1	17	6
Child 2	11	2
Child 3	15	12
Child 4	4	5
Child 5	14	11
Child 6	11	11
Child 7	7	10
Child 8	13	15
Child 9	14	6
Child 10	16	7
Child 11	10	7
Child 12	14	18
Child 13	8	16
Child 14	4	11
Child 15	15	18
Child 16	5	4
Child 17	17	15
Child 18	3	6
Child 19	18	16
Child 20	18	9



What is the relationship between the two tests? Is it strong, moderate, weak or non-existent? Why do you think this is the case?

You will find the answers to these exercises in the Resources section of the Real Training website. If you wish discuss your results with others, do so in the Course Forum.

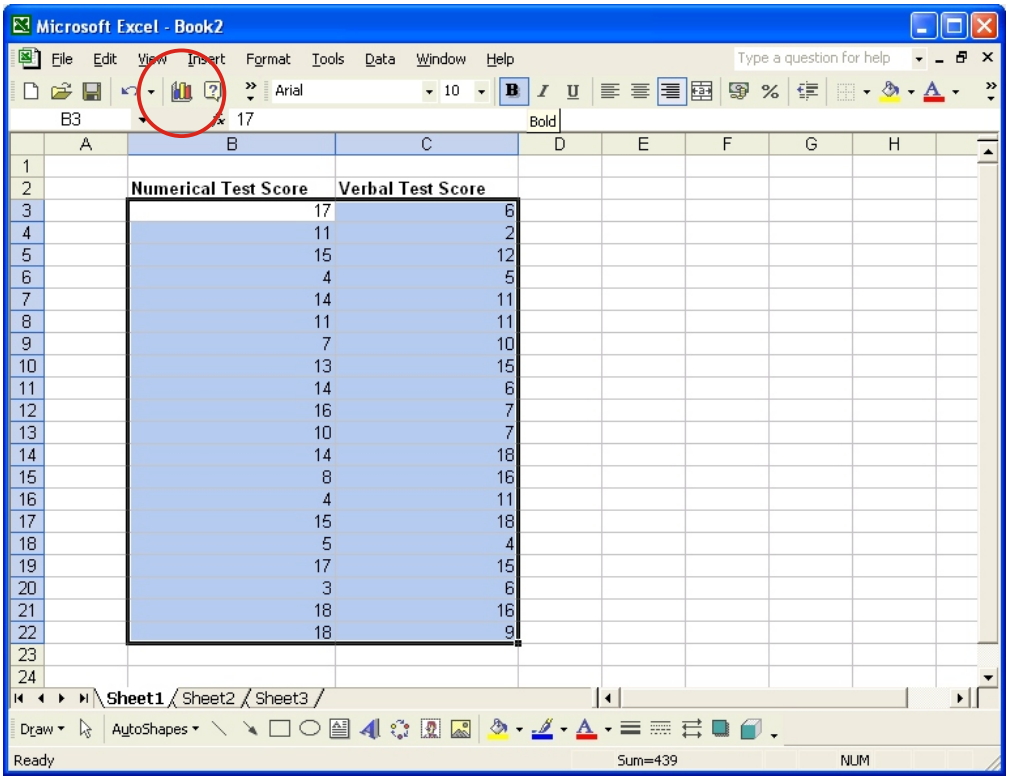
How to create a scattergram using Excel

Scattergrams can be easily created in Excel so, fortunately, if you have this program you won't have to do this by hand every time! Please note that it is sufficient for assessment of this competency that you are able to construct scattergrams by hand.

If you wish to learn how to use a computer to be able to do this first check to see if you have Microsoft Excel on your computer. It comes as part of Microsoft Office and if you have Word you are likely to have Excel. If you do not have it then you may be able to access it through your work. You may also wish to know that teachers, students and others working in education can often purchase Microsoft programs at a reduced price.

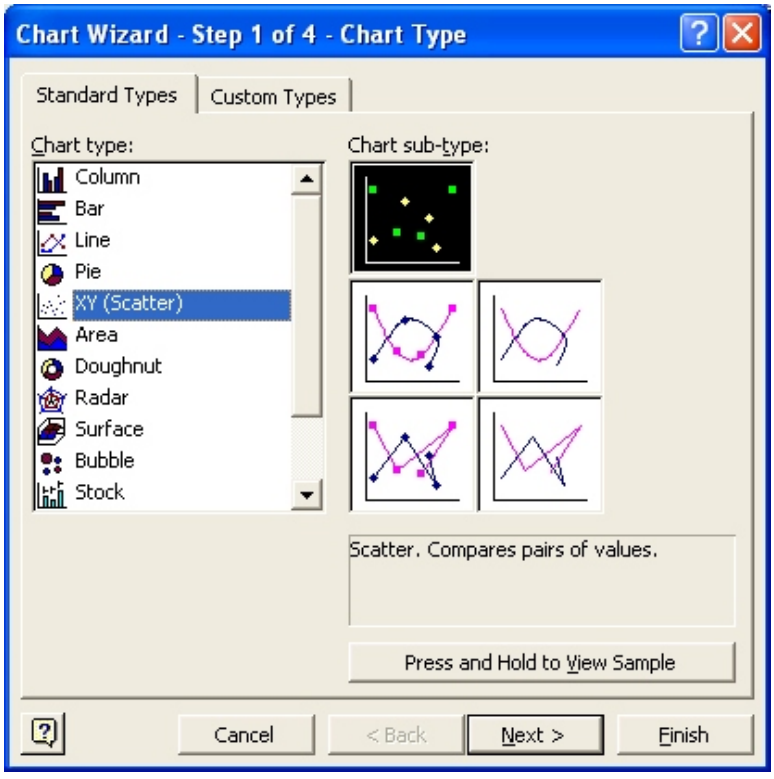
If you have Excel then all you have to do is type the data in so that you have two columns of values, the first column representing the first set of scores (i.e. from the numerical test), while the column on the right represents the second set of scores (i.e. from the verbal test).

Highlight all the scores and click on the Chart Wizard button (circled, it looks like a little bar chart), as shown in the picture below.

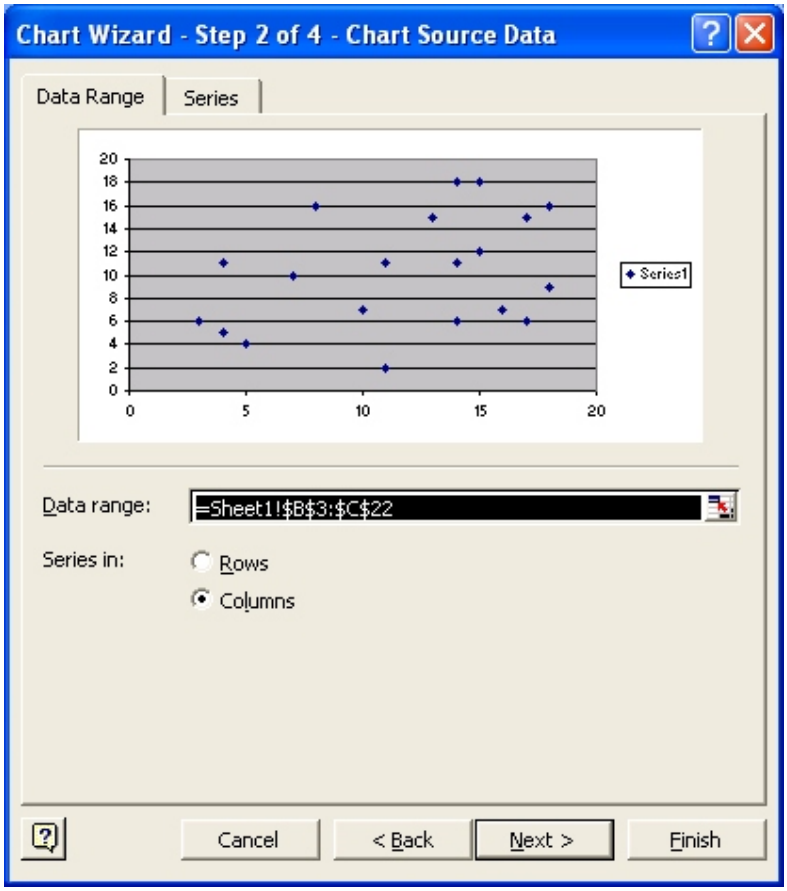


	Numerical Test Score	Verbal Test Score
3	17	6
4	11	2
5	15	12
6	4	5
7	14	11
8	11	11
9	7	10
10	13	15
11	14	6
12	16	7
13	10	7
14	14	18
15	8	16
16	4	11
17	15	18
18	5	4
19	17	15
20	3	6
21	18	16
22	18	9

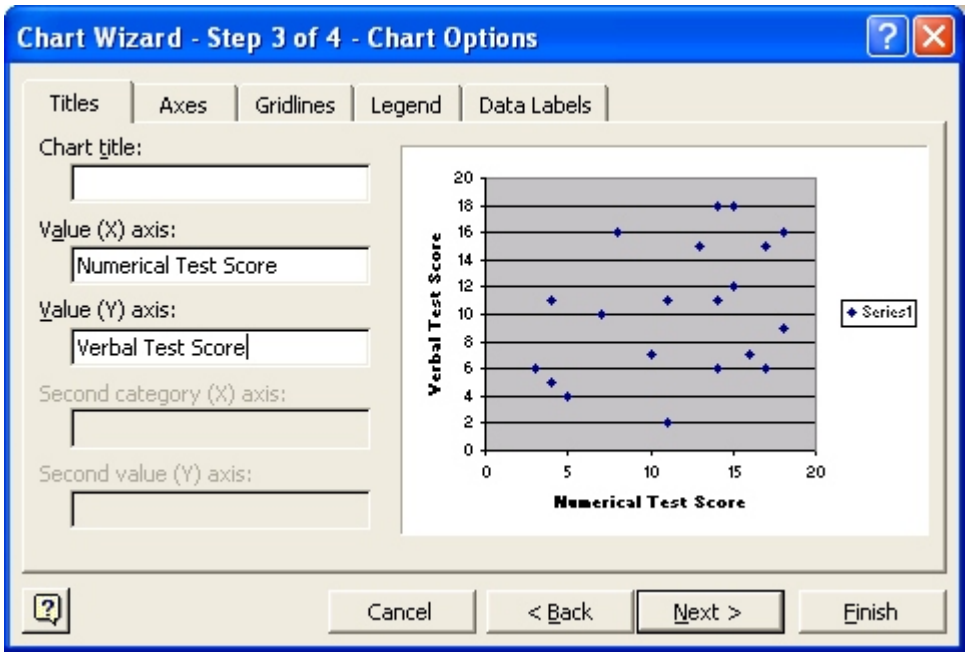
The Chart Wizard should start. In the first step, you need to select the type of chart you want. In our case, we want to select "XY (Scatter)". Click on your selection then click on Next.



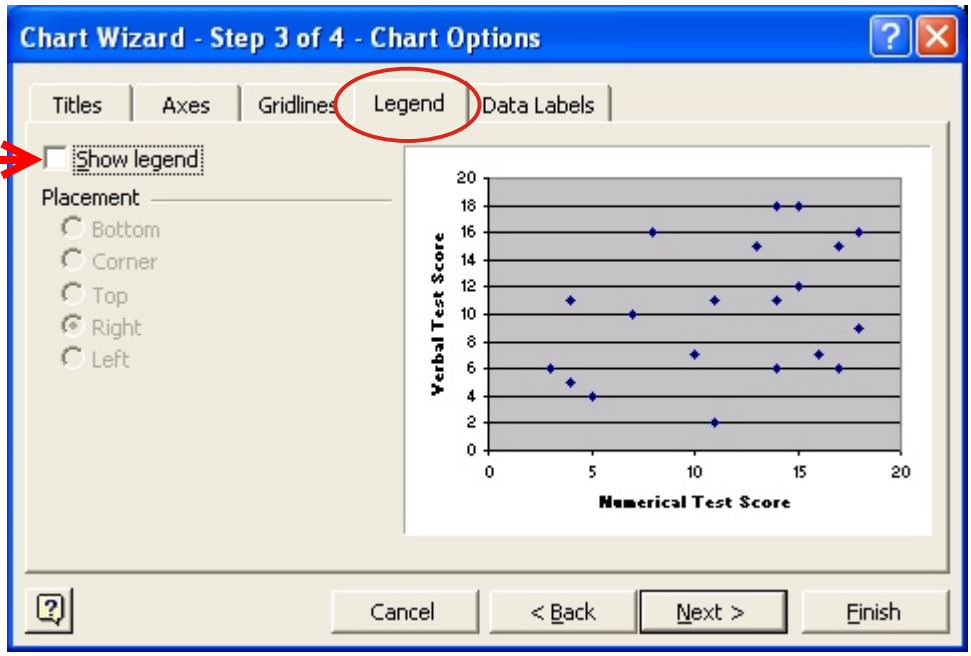
In Step 2, you just need to check that Excel has used the right data. From the display, it seems fine, so we just click on Next.



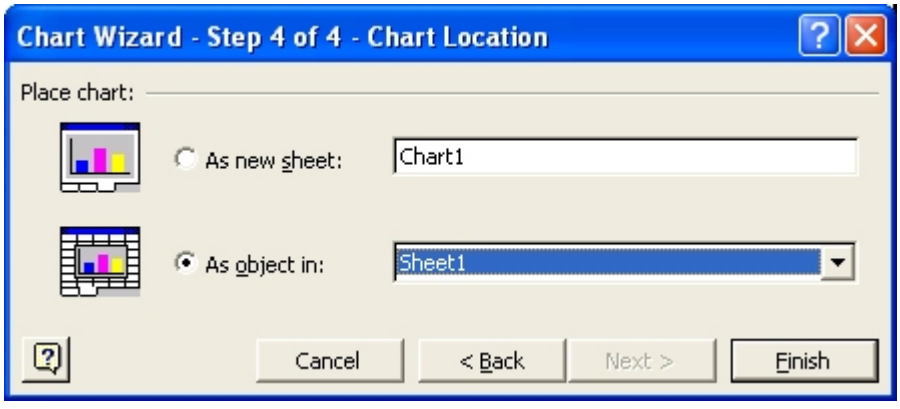
Then we get to Step 3 (nearly there!), where we specify our chart options. Here, you can add titles to your axes, so we know which scores are represented where. Simply type your information in the specified fields, as shown below.



Before we move onto Step 4, we can get rid of the legend, which we don't really need in this type of chart. To do this, click on the Legend tab and deselect the option "Show legend" by clicking on the ticked box. Click on Next when you're done.



In the final step, you simply need to tell Excel where you want your new chart. The second option (i.e. "As object in Sheet 1") is always better, because it makes it easier for you to modify the appearance of the chart.



Once you get your scattergram, play around with it to change its appearance. You can click and drag the whole thing to change its size. You can also modify virtually every aspect of the chart by double-clicking on the element you want to change.

Task 29 (optional)

If you have access to excel try this using the same data. Check your graph you drew against the one above.

If you want to, try using a different set of data that you have made up or better still take a sample of children/young people and administer two tests. Record the results and create the scattergram.

Remember to save this Excel file, it can be used later.

Go back to the scattergram you drew earlier. Do they look the same? If there are some differences, check to see what it is that you got wrong and make sure you understand the right way of plotting the dots onto the chart. Ask questions on the Course Forum if you need to or log onto a live chat.

Correlation coefficients

Creating a scattergram every time we want to show the extent to which two variables or tests are correlated would be a bit tiresome. That's why we have a numerical way of expressing the strength of this relationship: the correlation coefficient.

A correlation coefficient can range from -1.0 through 0.0 to +1.0. A correlation coefficient of 0.0 indicates there is no relationship whatsoever between the two variables, that is, they are fully independent. A correlation coefficient with an absolute value[1] of 1.0 is associated with a perfect correlation. In a perfect correlation, all the dots fall along a straight line (see Figure 3.3).

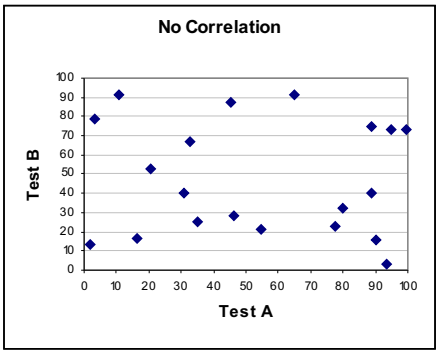
Pearson's r is the correlation coefficient that is usually given when dealing with continuous data (i.e. your scores are interval or ratio). For ordinal (ranked) data, **Spearman's rho** and **Kendall's tau** are used instead. You can find further information about interval, ratio and ordinal data on page 17 of A Psychometrics Primer by Paul Kline.

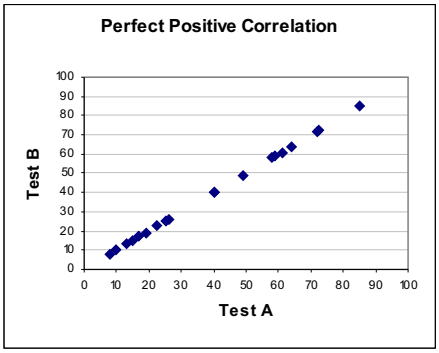
As can be seen in Figure 3.3(b), two tests are positively correlated when an increase in score in one test is associated with an increase in score in the other. In a negative correlation the situation is the reverse, that is, increasing scores on one test are associated with *decreasing* scores on the other (i.e. the higher you score in Test A, the lower your score will be for Test B). This might be the case where one test gave high scores for inattention and another test gave high scores for mental arithmetic attainment. Someone with low levels of attention (a high inattention score) might be expected to score low on a test of mental arithmetic.

For you to get a perfect positive correlation, you will either have a test being correlated with itself (i.e. you are correlating exactly the same set of scores) or have a constant relationship between the sets of scores (e.g. a score on Test B will always equal the score on Test A plus a constant number such as 10).

[1] An absolute value is the value of a number without taking into account its sign (i.e. whether it is positive or negative), so, for example, the absolute value of -1 is 1.

In the case of a perfect negative correlation, Figure 3.3(c), your scores on Test B will be the inverse[2] of those in Test A, or the inverse plus a constant.





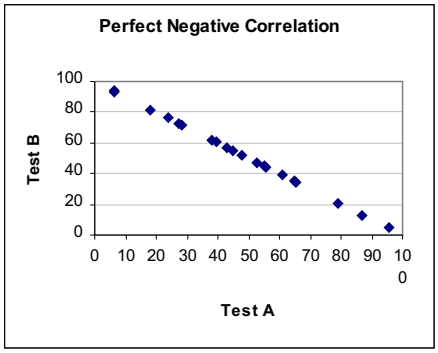


Figure 3.3 Scattergrams showing
(a) no correlation (r=0.0),
(b) a perfect positive correlation (r=1.0) and
(c) a perfect negative correlation (r=-1.0).

In real life, where we do not live in a perfectly predictable world, these extremely high correlation coefficients rarely exist and you will be more likely to come across something like what is shown in Figure 3.4 below. As you can see, both the positive and the negative correlations are strong. You determine the strength of a correlation by looking at the correlation coefficient. The higher the correlation coefficient (i.e. the closer in absolute value to 1.0), the stronger the correlation. In general, the continuum is used as follows (please note this applies to both positive and negative correlations):

[2]The inverse of a given score can be calculated by taking this number from the maximum possible score. For example, if we had a score of 46 out of a possible 100, the inverse of 46 will be 54 (i.e. 100 minus 46).

0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
no correlation	weak correlation			moderate correlation			strong correlation			perfect correlation

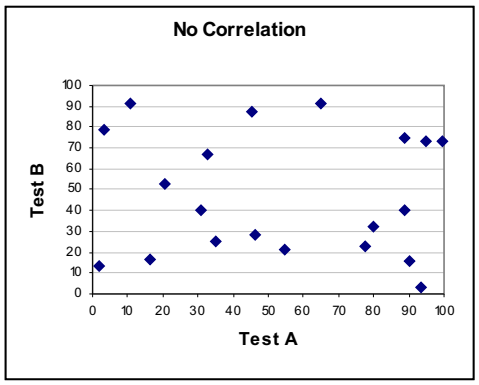
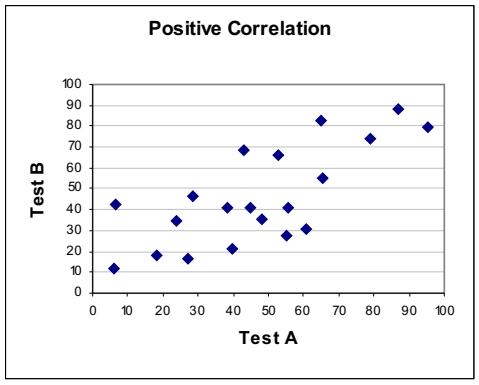
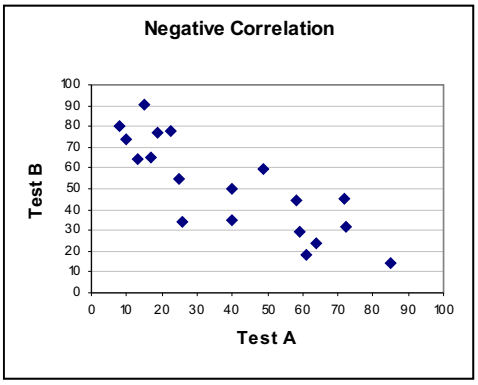


Figure 3.4. Scattergrams showing
(a) no correlation (r=0.0).
(b) a strong positive correlation (r=0.7)
(c) a strong negative correlation (r=-0.8)



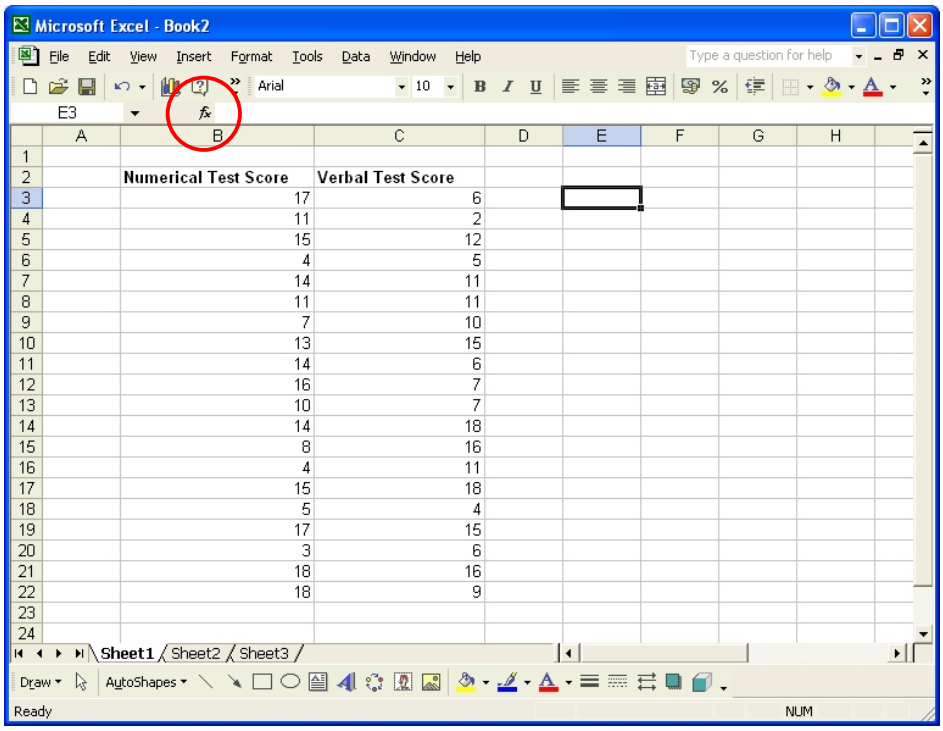


Reputable test publishers provide a range of information about reliability and validity and these are usually expressed or interpreted as correlation coefficients.You can now begin to see how using the guide above you will be able to judge whether the score provided is a moderate score or a good score. In general a "good test" will have reliability figures of 0.7 or above.We will discuss this further later.

How to calculate correlation coefficients in Excel

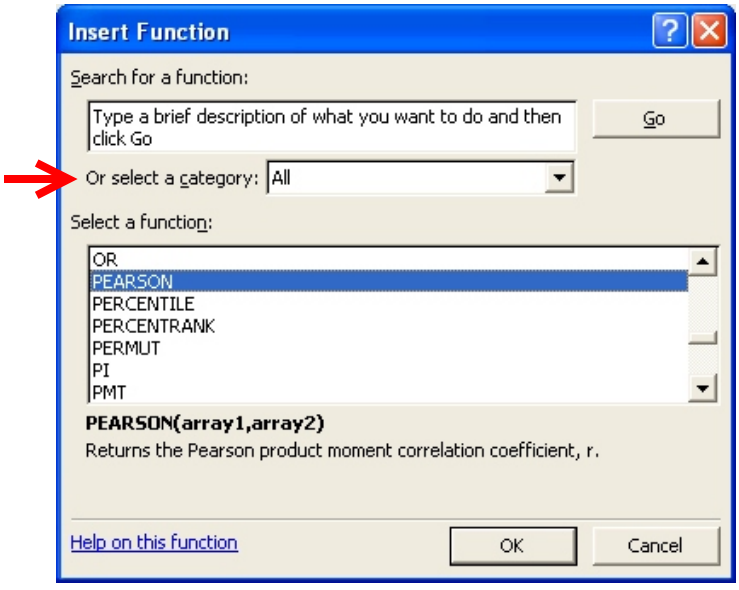
Excel can also calculate correlation coefficients for you. If you did it go back to the Excel file you saved earlier. Remember we had 20 children for whom we had scores on a numerical and a verbal test.

All you have to do is click anywhere on your worksheet and click on the function button (f_x, circled, if you can't see the button, click on Insert, then Function).



	Numerical Test Score	Verbal Test Score
3	17	6
4	11	2
5	15	12
6	4	5
7	14	11
8	11	11
9	7	10
10	13	15
11	14	6
12	16	7
13	10	7
14	14	18
15	8	16
16	4	11
17	15	18
18	5	4
19	17	15
20	3	6
21	18	16
22	18	9

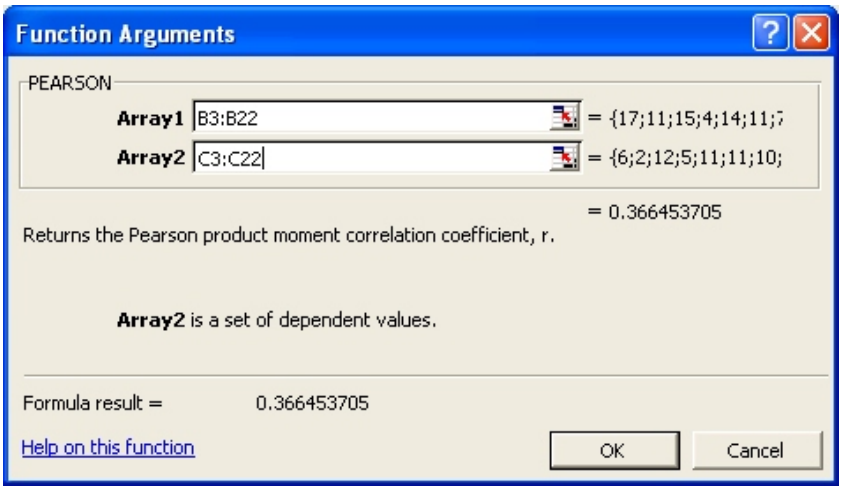
Select ALL in the "Or select a category:" drop-down menu and scroll down until you find "PEARSON". Click on this then click on OK.



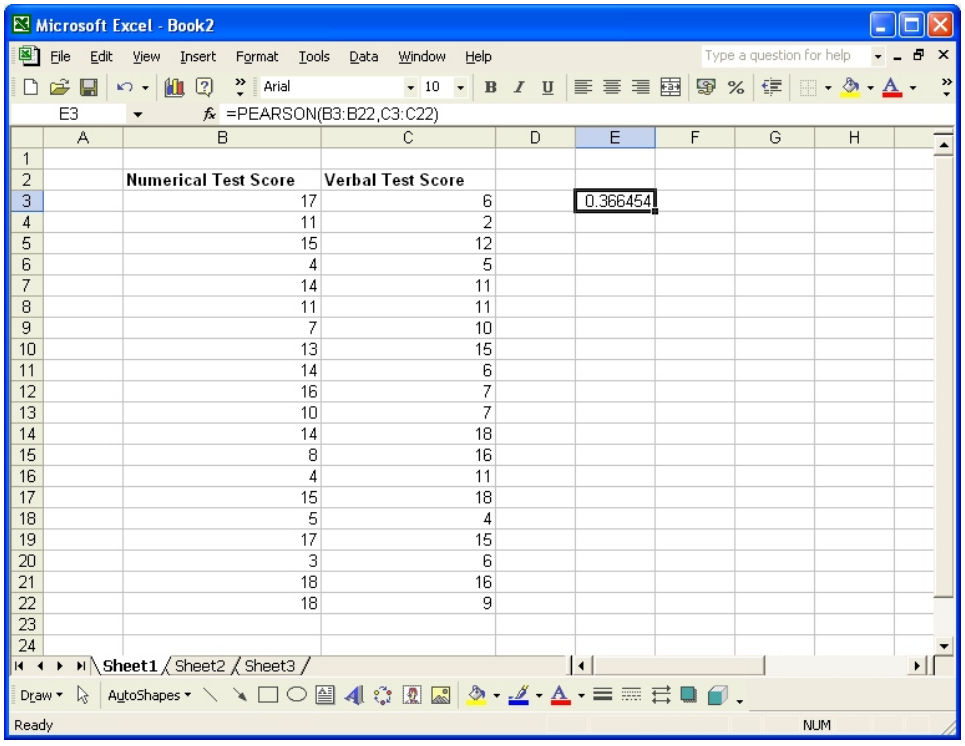
You will get the Function Arguments window. What you need to do here is to tell Excel what sets of scores you want to correlate. The first set (i.e. numerical test scores) will go into the Array 1 box while the second set (i.e. verbal test scores) will go into the Array 2 box.

You can enter these values by typing in the range, that is, the first cell followed by a colon then the last cell in your set. For example, in the case of the numerical test scores, the first score is located on cell B3 within the worksheet while the last score is located on cell B22.

An easier way of doing this is to position your cursor on the appropriate box (i.e. Array 1 or Array 2) then select the group of scores in that list. To do this put the curser over the first score in the list, left click the mouse and HOLD IT DOWN. Move the mouse to the last score on the list and release the left click. The list will automatically appear in the predetermined array. Repeat for array 2. When you're done, click on OK.



Your correlation coefficient should appear in the cell you were positioned in earlier, as shown below. You should have obtained a Pearson's r value of 0.37. This indicates there is a weak (nearly moderate) positive correlation between the numerical and verbal test scores.



TASK 30 (optional)

If you have access to excel try this using the same data. Check your co-efficient against the one above.

If you want to, try using a different set of data that you have made up or better still take a sample of children/young people and administer two tests. Record the results and then calculate the correlation co-efficient.

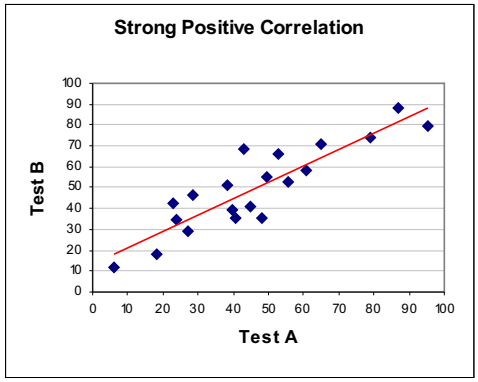
Summary

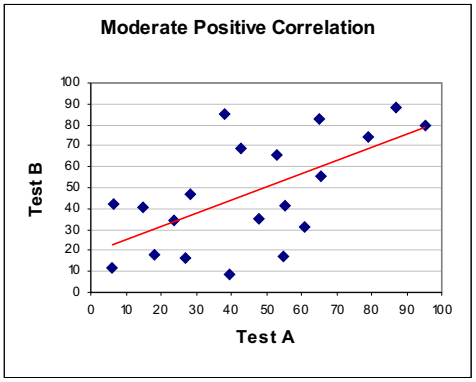
- Correlation coefficients are numerical indices of the relationships between two variables.
- A correlation coefficient of 0.0 indicates that there is no relationship between the two variables and that the variables are independent. Conversely, a correlation of +1.0 or -1.0 represents a perfect relationship between two variables.
- A positive correlation indicates that high scores on one variable are associated with high scores on the other variable.
- A negative correlation indicates that high scores on one variable are associated with low scores on the other variable.
- Pearson's Product Moment Correlation Coefficient (r) is the most commonly used measure of correlation for interval and ratio data. Spearman's rho and Kendall's tau are commonly used with ordinal (ranked) data.
- When we have a significant correlation, we can calculate the amount of shared variance between the variables by squaring the correlation coefficient.
- Correlations are used to estimate the reliability of a test. There will be more detail about this in competency 3.4.

Maximising and Minimising Correlations

It is useful to understand the conditions when correlations are more likely to be found (i.e. when they are maximised) and when they might not be found (i.e. when minimised). This can help you understand the circumstances when there really may be a correlation but it has not been found statistically. It can also help you to understand how reliability issues and measurement error effect the correlation coefficients provided as measures of reliability.

Correlations are maximised (both positively and negatively) when there is a very high degree of association between two variables AND there are scores available over a wide range. A high degree of association between two variables is apparent from a correlation coefficient high in its absolute value (i.e. close to +1.0 or -1.0). When looking at a scattergram, a high degree of association can be seen when dots are close to the trend line[3] (remember: dots along a straight line represented a perfect correlation of r=+1.0 or r+-1.0). For example, as can be seen in Figure 3.5, the closer the dots are to the line, the stronger the correlation.





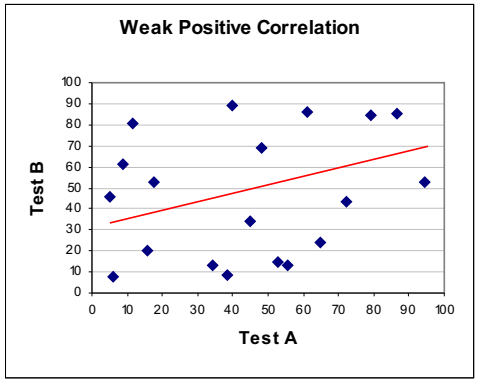


Figure 3.5 Scattergrams showing
(a) a strong positive correlation (r=0.9)
(b) a moderate positive correlation (r=0.6)
(c) a weak positive correlation (r=0.3).

Correlations are minimised if the range of scores is restricted or truncated. In the most extreme case, for example, the correlation between reading performance and gender would be 0.0 if the sample was made up only of boys, as gender would then be a constant (i.e. taking the same value each time), not a variable (i.e. there will be no variation in scores). In another example if you were seeking to calculate a correlation with some test data but only had test scores

[3] A trend line in a scattergram is a line that minimises the distance from each data point to such line.

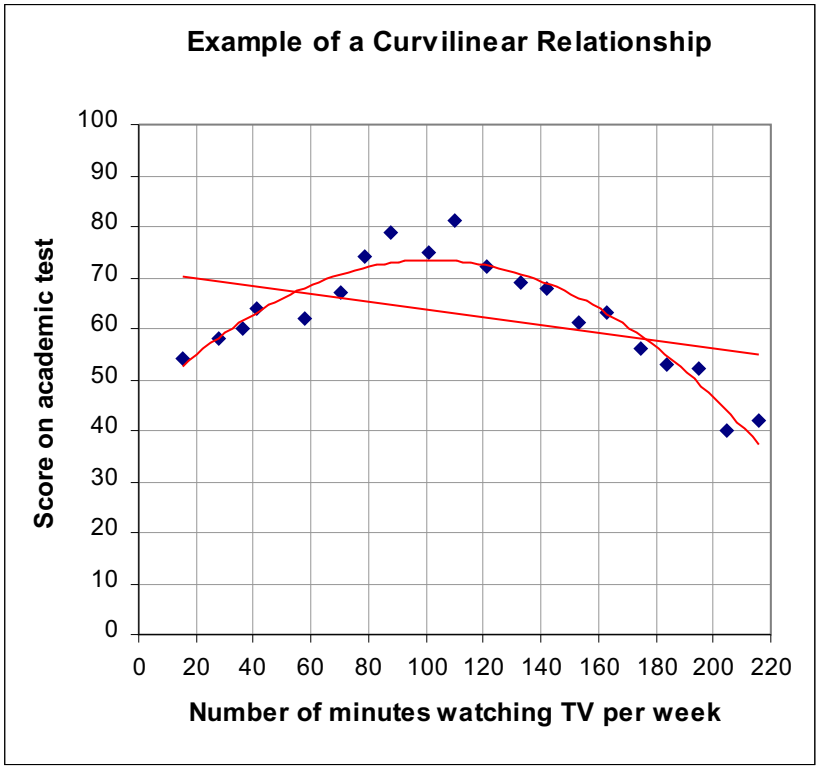
from three test takers and those test scores were very similar to each other the correlation would be more like to be maximised if you had data from more test takers across a wider range.

Note also that a linear correlation can be minimised if there is an underlying non-linear (i.e. curvilinear) relationship between the two variables. An example of a curvilinear relationship could be that between the amount of TV viewing and academic performance. In this case, you might get a situation where children who watch a moderate amount of TV perform best, while those who watch too little or too much perform a bit worse (e.g. because they are not benefiting from the educational aspect of TV or they are left with no time to do their homework, respectively). In this case, there is not a linear relationship, but a curvilinear one, that is, as the amount of TV watching increases, children's academic performance also increases until it gets to a point when they start to go down (see Figure 3.6).

If we were to calculate a Pearson's r correlation coefficient for these data, we would get a Pearson's r value of -0.43, which indicates a moderate linear negative correlation. If we could calculate a correlation coefficient that was designed to detect curvilinear relationships, we will get a much higher coefficient, since the data points (i.e. dots in the scattergram) will be much closer to the curved than the straight line.

Figure 3.6

Example of a curvilinear relationship between two variables (i.e. amount of TV viewing and scores on an academic performance test). Linear and curvilinear trend lines are also given.



TASK 31

Describe in your own words the conditions required for a high (maximised) positive correlation and a highly negative correlation. Write this below

If you want to discuss this further then use the Course Forum OR request a chat seminar on this topic by sending a message to your tutor. We will be very pleased to host an online chat seminar about specific topics.

Estimating Correlations

With all the information you have learnt so far, you should be able to estimate fairly accurate correlation coefficients from looking at a scattergram.

Remember that:

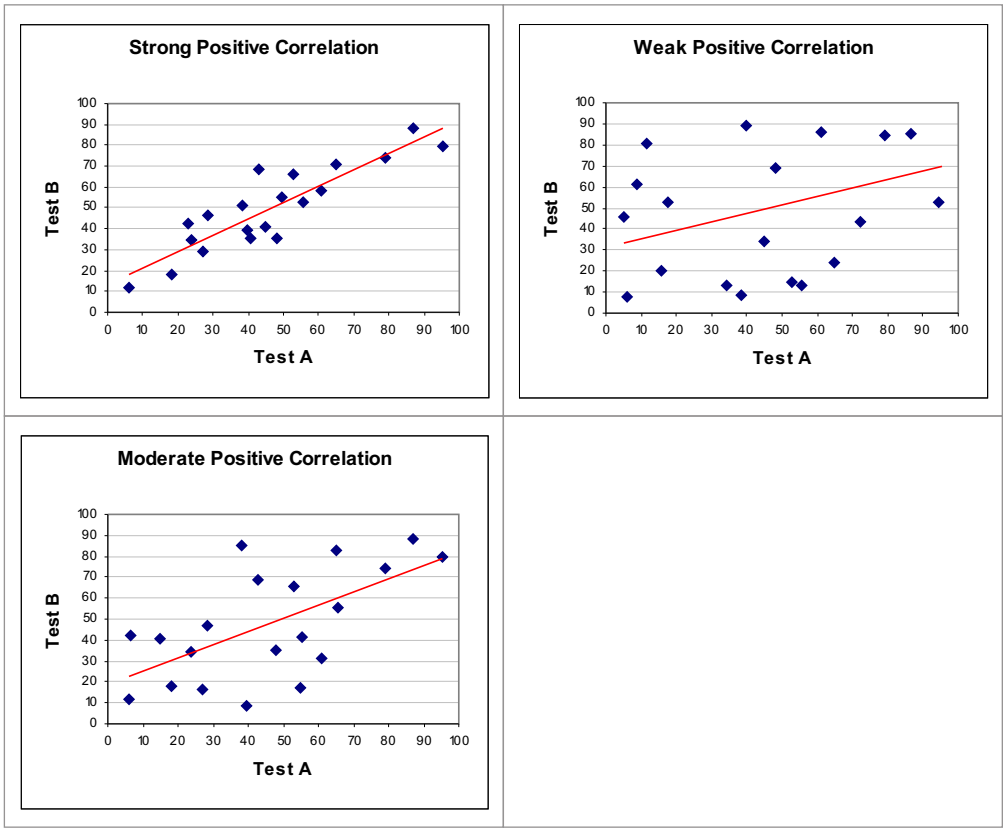
- if all points on the scattergram fall directly on a straight line such that high scores on the first variable are associated with high scores on the second variable, this represents a perfect positive correlation (r=1.0);
- if all points on the scattergram fall directly on a straight line such that high scores on the first variable are associated with low scores on the second variable, this represents a perfect negative correlation (r=-1.0);
- if data points do not fall directly on a straight line, you should create an imaginary line to represent the relationship between the variables (i.e. this should be a line that minimises the distance from each data point to such line) the closer the points are to the line, the stronger the correlation;
- if the range of scores in either of your variables is small, this will make your correlation coefficient weaker; and
- if the data points are more representative of a curvilinear rather than a linear association, this will also weaken your correlation coefficient.

You should also note that a correlation of 0.7 between two variables does not necessarily mean there is a strong causal relationship between these variables. For example, we might find a strong correlation between the number of hours spent with your friends outside and reading age in children. In this case, age will be a mediating factor: because age is highly correlated with both variables (i.e. the older you are, the more you're going to hang out with your friends outside and also the higher your reading age will be). This results in these two variables being highly correlated between them when, in fact, if the effect of age was to be taken out, the variables may not correlate at all.

TASK 32

It is very useful to be able to approximate the Pearson r value for different strengths of correlation. This will help you visualise and/or evaluate how strong a correlation is from the Pearson r value presented in a test manual.

Without looking back - look again at the graphs below. Can you remember or estimate the Pearson r value for each of the correlations represented in the scattergrams below? Remember Pearson r values range from zero (no correlation) to 1 (a perfect correlation) and in real life it is very unusual to get a figure higher than 0.9 for a very strong correlation.



If you want to discuss this further then use the Course Forum OR request a chat seminar on this topic by sending a message to your tutor. We will be very pleased to host an online chat seminar about specific topics.

Methods for estimating reliability

There are many types of reliability and therefore different methods of measuring or estimating reliability. Most of these, however, are based on correlations, which is why we learnt about correlation coefficients earlier.

Reliability measures relate to either the *stability* (how consistent or repeatable) or the *homogeneity* (internal consistency) of the test. First we will deal with measuring or estimating the stability of a test.

Measures of stability

Measures of stability refer to how consistent and repeatable the test is. These include:

Inter-rater reliability

The first measure of stability is inter-rater reliability. It is important to know that a test is not sensitive to who administers it. A good test would give the same score if a person was tested by different test administrators. You want to be confident that any difference in score is due to the skills or abilities of the person being assessed and not the error associated with test administration. Some tests are so difficult to administer or score (it is sometimes impossible to understand the instructions!) that different people administer it differently. These are the type of issues which cause low inter-rater reliability scores.

To calculate inter-rater reliability, you need two different people administering and scoring the test independently. A correlation between their total scores will give you the inter-rater reliability coefficient[4]. You now know how to calculate correlation coefficients and so you could work out an inter-rater reliability figure of a test used often in your school. As we discussed previously a score of 0.7 or above is a good score. Usually we would expect a score of 0.9 or above for this type of reliability.

When the range of possible total scores for a test is less than 15, it is more advisable to calculate the intraclass correlation coefficient (ICC), which reflects the degree of agreement (not correlation) between the raters (i.e. percentage of point-to-point agreements). To illustrate the difference between agreement and correlation, we could look at a scenario where Rater A gave scores that were consistently 20 points higher than Rater B. If a correlation coefficient was to be calculated, this will result in a perfect positive correlation. If an ICC was calculated, this would be zero, because the raters did not give the same score for any of the people being tested.

ICCs are also recommended when there are more than two sets of scores to be correlated.

Test-retest reliability

While inter-rater reliability assesses the consistency of a test across raters, test-retest reliability assesses the stability of a test across time. To calculate this correlation, a test is administered to a group of participants, and then re-administered again later to the same group. The scores from the two administrations are correlated to obtain the test-retest reliability coefficient.

[4]Remember that, when calculating the total score of a test, the scoring of reversed items needs to be taken into account. For example, if you had a five-point Likert scale (i.e. one to five) a score of '2' for the item "I enjoy reading" will still count as '2', while a score of '2' for the item "Reading is boring" will count as '4' (i.e. you reverse the scale).

Guidelines regarding the time period between assessments are not clear, with suggestions for minimum periods ranging from two weeks to three months (e.g. Kline, 2000). The problem is that, the greater the time period, the more likely the scores will change and therefore the lower the test-retest reliability. However, if the time period is too small, practice effects might interfere with the results.

A test-retest reliability coefficient of at least 0.7 would be acceptable for a published test, however the higher the score the better. The threshold is higher (0.9) for ability tests which, by definition, should be more consistent than tests measuring other, more naturally changing psychological variables (e.g. anxiety).

Because there are different ways to set up this mini test-retest reliability data gathering the test manual should also describe how the test-retest reliability data was collected.

Test-retest is an important concept in the construction of tests. You need to be sure that the test will measure in the same way whenever you use it. You won't have time during this unit but if you are interested you could set yourself up with a small test-retest experiment which would be completed later.

Choose a test that you know well and administer it to a reasonably large sample. We recommend you use a group test so that you don't have to do it individually that would simply take too long. Then retest the same children and calculate the correlation co-efficient. The time interval between testing sessions should be long enough to minimise practice effects and short enough to minimise variables open to change (e.g. learning). Does your result compare with that published in the test manual? If not, what could account for the difference?

Note this is not a course requirement but if you are by now interested in the way in which tests are constructed this might be worthwhile. You will certainly gain a deeper insight into tests and their reliability.

Measures of homogeneity

For a test to be reliable, it doesn't only need to give consistent and repeatable scores (i.e. across raters, time, etc.), but also to be homogeneous, that is, consistent *within itself*.

Measures of homogeneity therefore relate to the internal consistency of a test. In other words, are all the test items measuring the same construct? If for example the test is measuring sequencing skills then half of the items should correlate with the other items as they are all measuring the same thing! There are two main approaches to this Same/alternate form and split-half reliability.

Same / alternate form reliability

In this case, two equivalent forms of a test are developed (i.e. Test A1 and Test A2 this is common for reading tests like the Neale Analysis) and these are administered to a group of participants in counterbalanced order[5]. The scores of these two alternate tests are correlated to obtain the same/alternate form reliability coefficient.

[5]To counterbalance the order of administration, you would get half of your participants to do Test A1 first, then Test A2, and the other half to do it in the reverse order (i.e. Test A2 first, then Test A1).

A coefficient of at least 0.9 is usually required and this can decrease the longer the lapse between the two administrations. Again, we have the same problems as with test-retest reliability, since too small a time interval can lead to confounded scores (i.e. affected by fatigue, or practice effects for example).

In practice, this form of reliability is hardly used because it's just too much work since you need to create twice the amount of items and also make sure that the two tests are equivalent, which can be extremely difficult.

Split-half reliability

This second approach is much more widely used because only one version of the test is required. In this case, the test is administered to a group of participants and then items are randomly split into two groups (e.g. even- vs. odd-numbered items). A total score for each of these groups of items is calculated and the correlation between these will give you the split-half reliability coefficient. It is like splitting the test in half and treating each half as if they are the same or alternate forms. They should be giving scores that correlate strongly with each other.

There are many variations to this technique. For example, you might want to correlate each individual item score with the total score for the test then calculate the average of these correlations (i.e. 'item-to-total correlation coefficient). The most commonly used measure is Cronbach's (1951) alpha coefficient, which provides an average split-half correlation coefficient based on all possible divisions of a test into two parts. Richardson's coefficient is used when the answers to the items are dichotomous (e.g. yes/no) instead of being in Likert scale form.

Some authors (e.g. Cattell, 1957) have suggested that a test could be *too* internally consistent. Their argument is that, if the items of a test are very highly correlated with one another (i.e. 0.8 and over), this could indicate that some of the items are redundant. This might well be the case, especially for tests with a small number of items.

It should be clear by now why the concept of correlation was so important before we looked at how the reliability of a test is estimated. These general guidelines apply:

- a reliability coefficient of 0.9 and above indicates a high degree of reliability (81% consistent variation (or 81% of the variance is accounted for), 19% error);
- a reliability coefficient between 0.8 and 0.9 indicates a moderate degree of reliability (64% consistent variation, 36% error);
- a reliability coefficient between 0.7 and 0.8 indicates a low degree of reliability (49% consistent variation, 51% error); and
- a reliability coefficient of less than 0.7 is unacceptable.

TASK 33

In your own words, write a brief explanation of the following forms of reliability and identify their relative pros and cons (you may also want to have a look at chapter 3 A Psychometric Primer by Paul Kline). **You can also type "define: test-retest reliability" and "define: inter-rater reliability", without the inverted commas, into to website www.google.com for further definitions.**

inter-rater reliability

test-retest reliability

same/alternate form reliability

split-half reliability

If you want to discuss this topic further then use the Course Forum OR request a chat seminar on this topic by sending a message to your tutor. We will be very pleased to host an online chat seminar about reliability.

TASK 34

Consider the following scenario:

Gail has been using the same reading test for years. Someone else has recently used the same test with the same young person (or adult) just a week after she had tested their reading. It was a mistake and the young person (or adult) did not even say they had completed the same test the week before! What was interesting was that the scores came out different. For some reason or reasons, Gail's score gave a reading age equivalence of 6 months below the score that her colleague found 1 week later. How could this happen? Think of as many reasons as possible.

Type your thoughts in the appropriate section of the Course Forum on the Real Training website. Compare your thoughts with at least one other and reflect on whether you feel you can explain these concepts. Reflect on this exercise and make some notes below.

Remember to print out and save your contributions to the Forum in your Real Training folder as part of your portfolio of evidence.

Methods for estimating validity

Methods for estimating validity are not as straight forward as those for reliability. For most types of validity, there are no numerical measurements, which makes this process much more subjective, as you will see.

As with reliability, there are different forms of validity, including:

Face validity

A test has face validity if it appears to measure what it claims to measure. It is mainly about the items in the test looking relevant to the participants. Although it is not a great measure of validity, it helps by not discouraging participants from completing the test (i.e. if the items look irrelevant, a participant might be put off).

Content validity

Content validity refers to the extent to which the items in your test are a representative sample of the universe of items for the relevant variable. In other words, if we have a variable to be measured, in theory at least, there must be a universe of items that can be used to measure such variable. For a test to have content validity, its items need to be an accurate representation of this universe of items.

This requires that only one variable is measured by your test and that a clear conceptualisation of your variable is available.

Construct validity

Construct validity is similar to content validity, in that it requires that all theoretical dimensions of the variable are covered in your test. It requires for the variable to be operationally defined. An operational definition of a variable breaks the concepts into smaller, more specific aspects (or behaviours) that can be scientifically measured.

For example, if you wanted to measure depression, you need to decide what behaviours you would be looking for. These may include 'feeling hopeless', 'loss of appetite', etc. These specific behaviours or emotions should span across all the dimensions of the concept. In our case, 'feeling sad' would tap into the emotional dimension, while 'loss of appetite' refers to the physiological symptoms associated with depression.

The test manual should include theoretical information about the content and construct being measured. There should be research information about this area. There should then be an analysis of each item and how each item relates to the theory. Ideally this analysis should have been undertaken by an acknowledged independent expert in the field. For example the Wechsler Objective Reading Dimensions reading comprehension subtest includes an analysis of the different type of comprehension skills (e.g. factual recall, inference) being assessed by each individual item.

Criterion-related validity

In criterion-related validity, you look at the relationship between a new test and a relevant criterion. This criterion could be an already well-established test (*concurrent validity*) or a variable that is expected to be predicted from our test (*predictive validity*).

In the case of concurrent validity, the scores of your new test are correlated with those of the test that is well established. You might wonder why we would want to develop a new test if there is already a perfectly useful one that

measures the same variable. In fact, there are many reasons for this. It could be that the established test is too long, or only suitable for group administration. A new test might be more practical and suited for different circumstances.

So a new reading test's validity is often checked against well established and respected reading tests thus borrowing from some of the old tests credibility.

Predictive validity requires us to identify a suitable variable that has been demonstrated to be highly correlated with the variable we want to measure. Ability test, in particular, are designed to be predictive. For instance, if you were to develop a new intelligence test, you will want this to be predictive of whether your participants will perform well academically.

The problem with criterion-related validity is that a suitable criterion needs to be identified and this can be difficult. In those cases where a suitable criterion is not found, associated variables need to be used instead, but this might result in weaker correlation coefficients (e.g. 0.3 to 0.5).

Other factors are also important in making sure the test is measuring what it is supposed to. For example, wording is extremely important in ensuring accuracy in responses. Something like 'often' can be interpreted in different ways by different people, so it's better to use more objective phrases such as 'more than once a week'.

Transparency and social desirability (i.e. where the participant can guess the most desirable answers and gives these instead of an honest answer) also need to be considered when evaluating a test.

TASK 35

In your own words write a brief explanation of the following forms of validity and identify their relative pros and cons. You may want to have a look at chapter 3 of A Psychometrics Primer by Paul Kline. You can also type "define: face validity", "define: content validity", etc., without the inverted commas, into to web site www.google.com and you will find a range of definitions.

Face validity

Content validity

Construct validity

Concurrent validity

Predictive validity

If you want to discuss this topic further then use the Course Forum OR request a chat seminar on this topic by sending a message to your tutor. We will be very pleased to host an online chat seminar about validity.

TASK 36

Consider the following scenario:

Hassan has noticed that the reading test he uses seems to give good scores for young people (or adults) who still seem to have difficulties understanding written work in lessons (or lectures) He is starting to lose confidence that the test is providing a true measure of reading and in particular reading comprehension. How can he justify not using the test anymore and instead rely on classroom observations and individual interviews? Or maybe he could search for a better test. How could he know, before he used the test, whether it may be better at measuring reading comprehension skills?

Type your answer in the appropriate section of the Course Forum on the Real Training website. Compare your answer with at least one other and reflect on whether you feel you can explain these concepts. Reflect on this exercise and make some notes below.

Remember to print out and save your contributions to the Forum in your Real Training folder as part of your portfolio of evidence.

TASK 37

SUMMARY TASK FOR THIS UNIT

The aim of this task is for you to show you can justify your choices in selecting a good psychological test.

Think of a domain you want to measure (e.g. reading, a specific language skill) and a matching test either you own or may wish to use in the future. Have a look at the manual (or the test publisher's catalogue) for information about the reliability and validity of the test. Examine closely the information given regarding the reliability and validity of the test. Write some notes on how the reliability and validity has been investigated and reported. Critically analyse this information. If possible compare this test with another similar test and this will help you see which is potentially more reliable and valid.

Now prepare a paper of no more than 500 words entitled "Evaluating the reliability and validity of a psychological test of [whatever it is you want to measure]" (e.g. "Evaluating the reliability and validity of a psychological measure of spelling"). Upload it to the Real Training website. Upload your file to the Work Submissions area of the website. Do not use the Course Forum on the Real Training website. Help is available on the website or by e-mail at help@realtraining.co.uk. You should also save it in your Real Training folder as part of your portfolio.

Assessment of Unit 3

Assessment of the competencies in Unit 3 is by a straightforward questionnaire set to you shortly before a Competence Day. You may complete the questionnaire using your workbook and any books of papers that you wish. You will send us back you questionnaire before the Competence Day.

There will also be a few short exercises at the Competence Day to check your understanding of reliability coefficients.

References

Cattell, R.B. (1957). *Personality and Motivation Structure and Measurement*. Yonkers:World Book Co.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

Kline, P. (2000). *A Psychometrics Primer*. London: Free Association Books.

Trochim, W.M.K. (2002). *Reliability*. Last accessed 5th October 2004 from http://www.socialresearchmethods.net/kb/reliable.htm.

Further Reading

Buley, J. (2000). *Reliability, Validity and Correlation*. Last accessed 5th October 2004 from http://com.pp.asu.edu/classes/jerryb/rvc.html.

Kline, P. (2000). *A Psychometrics Primer*. London: Free Association Books.

Palmquist, M. (2004). *Reliability and Validity*. Last accessed 5th October 2004 from http://writing.colostate.edu/references/research/relval/index.cfm

Trochim, W.M.K. (2002). *Reliability*. Last accessed 5th October 2004 from http://www.socialresearchmethods.net/kb/reliable.htm.
[Check other links within this website for more relevant material for example http://www.socialresearchmethods.net and http://www.socialresearchmethods.net/kb]

Young, F.W. (2004). *Correlation: The relationship of two variables*. Last accessed 5th October 2004 from http://forrest.psych.unc.edu/research/vista-frames/help/lecturenotes/lecture11/overview.html.