

The basic principles of scaling and standardisation

Introduction

This unit deals with the fundamental statistical concepts required to use educational tests.

If you are working in the field of assessment and want to explain clearly what tests actually mean this unit will provide you with the skills to do so. It will also allow you to compare one score in relation to another and make informed decisions about the severity of the difficulties, or magnitude of the strengths measured by a test. By the end of this course you will be able to clearly and confidently explain how percentiles and age equivalence scores are calculated and explain the results to children, parents, teachers and other adults.

In order to do this you will learn enough maths and statistical knowledge to allow you to understand what psychometrics and standardised testing is all about. This knowledge will also help you understand the test scores produced in reports by speech therapists, specialist teachers, psychologists and paediatricians.

You may or may not enjoy maths. That is not important. This unit will help you to be confident to interpret any normative test score. Learning this skill will be useful to you.

There is one key mathematical idea, or assumption that you need to learn. All standardised testing is based on the assumption that an individual's score can be compared to the scores of others who previously completed the test (the standardisation sample). You need to learn about the normal curve (also known as "the bell curve"). The key idea is that the scores of the children who previously completed the test can be graphed to show a normal curve shaped graph. You need to learn about how mathematicians, statisticians and psychometricians use percentiles and other standard scores to describe different positions on this normal curve.

Section 1

Competencies covered in this section of unit 2

Descriptive statistics

- 2.1 Construct a frequency distribution graph to illustrate how a sample of test scores accumulates at different points throughout the range covered by the test.
- 2.2 Undertake calculations to convey how much variation there is amongst a set of test scores.
- 2.3 Describe how the alternative ways of expressing the average of a set of test scores (mean, median or mode) are affected by their distribution throughout the range of scores.

Constructing
a frequency
distribution graph

Don't be put off by the terminology. You will have done this very same thing in primary school. You may have counted how many pupils had different colour eyes or perhaps you measured the height of everyone in your class.

Discrete distributions

In fact these two examples illustrate the two main types of frequency distribution. The first (eye colour) is an example of discrete distribution; that means that each classification is separate from each other classification. You make a decision as to whether the eyes are blue or brown or whatever other colours are being recorded and then add up the numbers in each group. Note it's not always easy to decide whether eyes are blue or brown. Error will creep into the results as different assessors might make different decisions!

Interval distributions

For the second type (e.g. height) you actually have to decide what range of heights to put in each group. This is called an interval distribution because the scores actually fall in a continuous range and to make some sense of it you need to decide how to group them. So in our example you might decide to use 5cm intervals and have groups for 1.40m-1.44m, 1.45m-1.49m, 1.50m-1.54m ... and so on.

Interval distributions are used more often in psychometric measurement because usually tests generate a range of scores out of a possible score. For example, a reading test has a finite number of items that can be scored and each person who does the test will get a number correct. Tests are constructed to place each score in one of the intervals or groups of scores so that their place in the range of scores can be identified. This is the position in relation to the normal distribution about which there will be more later. First let us return to first principles and start with a histogram. This is the most common and straightforward way to construct a frequency distribution of normal and non-normal scores.

40 people took a spelling test. There were 20 items on the test and the following scores were recorded.

1, 3, 4, 5, 5, 6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10,
11, 11, 11, 12, 12, 12, 13, 13, 13, 14, 14, 14, 15, 15, 16, 17, 18

If you chose an interval of three scores it gives you seven intervals. You could choose more intervals or less by changing the interval width. There are mathematical formulae for choosing the interval width but we recommend that you experiment and choose the width that best communicates the shape of the distribution. (For more information on this visit

<http://cnx.rice.edu/content/m10160/latest/>)

Seven intervals give you the following data table

Score interval	0-2	3-5	6-8	9-11	12-14	15-17	18-20
Number of scores	1	4	9	12	9	4	1

If you now make this into a histogram like the one below you can see the number of scores in each group graphically represented. Remember doing them at school now?

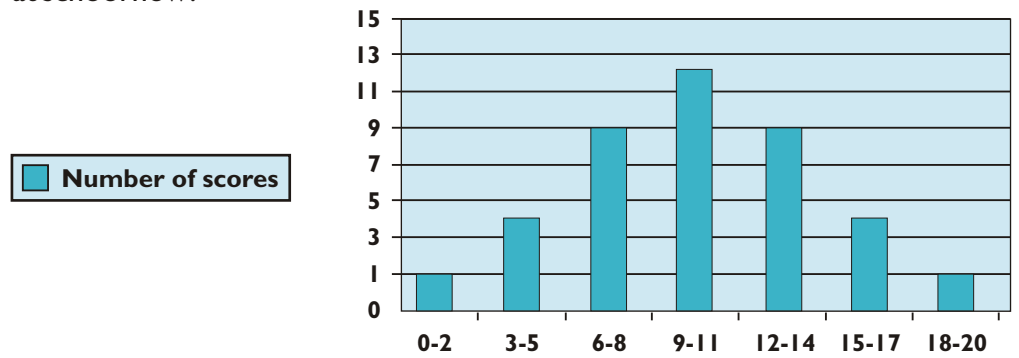


figure 2.1

TASK 10

Use the following data spelling the scores to draw your own histogram.

34, 32, 27, 26, 22, 20, 19, 18, 18, 16, 16, 14, 14, 14, 14, 14, 14, 13, 13, 13, 13, 12, 12, 12, 12, 12, 11, 11, 11, 11, 10, 10, 10, 8, 7, 7, 7, 6, 6, 6, 6, 5, 5, 4, 4, 4, 3, 3, 3, 2

First you will need to decide how to group the scores - what intervals to use.

You might choose intervals of 0-4, 5-9, 10-14 ... etc.

Complete the table below and draw your own histogram. Remember to label it.

Score interval							
Number of scores							



figure 2.2

You may find it useful to transfer this to graph paper as you can draw it to a size to suit you. You will use this data again so spend time understanding what you are doing with it. If you need to raise questions then use the Course Forum on the RealTraining website.

Variation in scores

You will notice just from looking at the scores above that they vary from a high to a low point and they vary from each other. Unless everyone scores the same this is axiomatic, in other words it is necessarily true and does not need to be proved. The chance of everyone scoring the same on a reasonable sample size is infinitesimally small and we do not need to concern ourselves with that possibility on this course. What is important is that the scores do vary and that you can work out how much they vary. The way we do this is first to establish the arithmetic *mean* (average) of the scores and then measure how much the scores vary from this average. Note, there are other types of average but we shall deal with them later.

How far the scores vary from the *mean* is called the variance, which is a numerical index that describes the dispersion or variability of a set of scores around the mean of the distribution. It is calculated by expressing the average squared distance of the scores from the mean.

Thus the formula for calculating the variance is:

$$\sigma^2 = \frac{\sum x^2}{N}$$

Where σ^2 is the symbol for variance, Σ is the symbol for adding up all the scores, x is the deviation of a score from the mean and N is the number of scores actually recorded. This looks more complicated than it is. In common language you calculate the distance each score is from the mean, you square all the calculations separately, add them all up and then divide by the number of scores you started with.

Although variance is important in psychometric theory it has more limited application in interpreting scores. More useful is the *standard deviation*, which is simply the square root of the *variance* and is expressed by the formula:

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

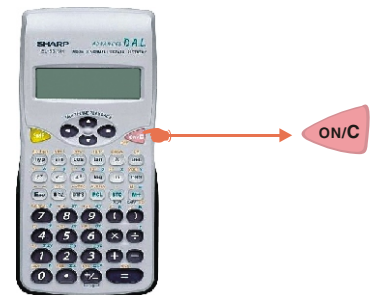
Note if σ^2 is the symbol for variance, if we drop the squared symbol (the superscript 2) then σ becomes the symbol for standard deviation, because σ squared is the variance.

It is therefore perfectly possible to work out the standard deviation of a set of scores using nothing more sophisticated than a paper and pencil. It takes some time, particularly with a large group of scores, but if you are so inclined it does help with understanding the process. However, if maths is not really your thing and you just want the quickest way to the answer then we have the solution.

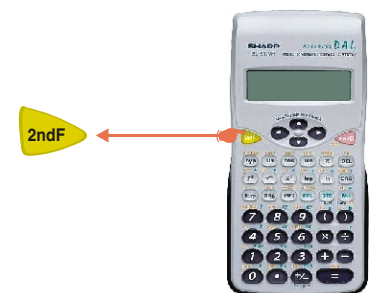
With your course materials we sent you a scientific calculator. You will probably never need the vast majority of its functions unless your job already involves a substantial amount of statistics in which case you probably already have one and the next few paragraphs will be teaching grandmothers to suck eggs. But if you are immediately turned into a cold sweat at the very thought of switching on such an instrument don't worry. It really is very easy and we will show you with simple steps how to not only calculate the standard deviation but also to check that you have got it right.

TASK 11 *Calculating the mean the easy way*

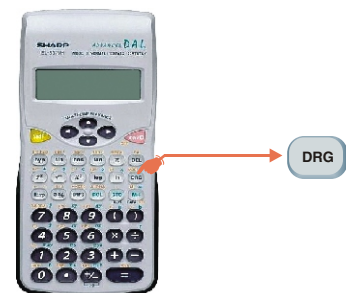
1. *Switch the calculator on*



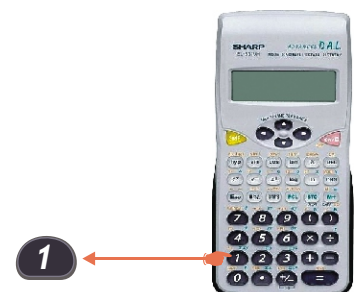
2. *Press the key marked 2ndF (second function)*



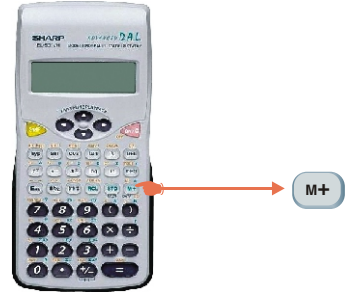
3. *Press the **MODE** key marked DRG on the key. **MODE** is above and to the left this is its second function)*



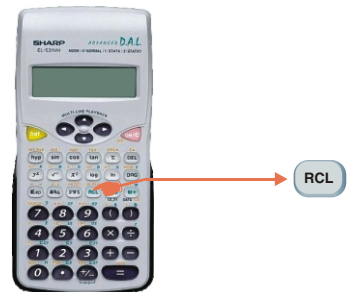
4. *Press 1 (for the Statistics mode. Note there is a reminder of the mode functions written on the calculator just above the display screen).*



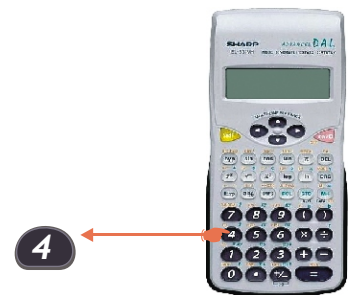
5. Enter each of the scores, one at a time, in the set of data you used for the previous task. After entering each one press the DATA key marked **M+** on the key (DATA is in black underneath). Note the calculator keeps a record of how many sets of data you have entered. If you make a mistake its best to clear the calculator and start again! To do this, see the note below. When you have finished n should equal 50.



6. Press the key marked RCL (ALPHA above this is its second function). You will see ALPHA appear at the top of the screen.



7. Press the key marked \bar{x} or 4 on the key.



8. The answer should appear with $\bar{x} =$.

Write down your answer



You have just found the answer to the arithmetic mean. This is calculated according to the formula below. It equals all of the numbers (x) added together (Σx) which are then divided by the number of pieces of data in the sample (N).

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

The first thing you should now do is make a judgment as to whether this makes sense. Is the answer even in the range of scores that you entered? If not it is wrong and you have made a calculation error. However more likely it is somewhere in the range and moreover somewhere near where you would judge the middle to be. In which case it is probably correct.

At this stage it is worth checking your use of the calculator by actually doing the calculation on paper. This might seem laborious and unnecessary but it is useful to give you confidence that the calculator does actually give the correct answer and it also provides you with an opportunity to understand the mathematical processes that underpin this statistical part of the course. But don't worry it really doesn't get any harder mathematically than this.

How to delete the data stored in the calculator

If you do not do either of the following your data will be preserved. As it happens this is useful as you will be using this data again. Just take care to remember which set of data you actually have in your calculator.

Note if you want to delete the data in your calculator you have to do one of the two following things:

Press the key marked 2ndF (second function). Then press the key marked DRG (MODE above and to the left - this is its second function). Then press 1 of the statistics mode or 0 for the normal mode.

Important note: You may need to turn the calculator back from statistics mode to the normal mode when carrying out normal calculator functions.

Or:

Press the key marked 2ndF (second function), followed by DEL (CA is its second function which means cancel all).

TASK 12

Calculating the mean the slightly harder way

On a piece of paper add together the scores you used previously.

34, 32, 27, 26, 22, 20, 19, 18, 18, 16, 16, 14, 14, 14, 14, 14, 14,
13, 13, 13, 13, 12, 12, 12, 12, 12, 11, 11, 11, 11, 10, 10, 10, 8,
7, 7, 7, 6, 6, 6, 6, 5, 5, 4, 4, 4, 3, 3, 2

Enter the total

Write down the number of scores in the list

Divide the total by the number of scores in the list and write your answer in the box

Is this answer the same as in TASK 11 above? It should be. If not go back and check both calculations. If you are still having trouble then log onto the Real Training website and seek some help through the Course Forum.

Calculating the standard deviation

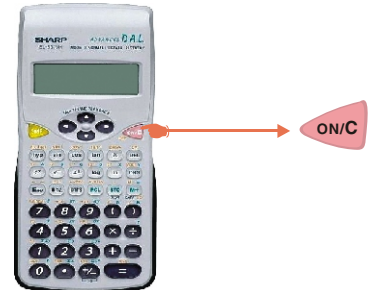
You are now ready to calculate the standard deviation. This can be done in two ways. Both involve a calculator because calculating the square root is difficult without one. The first method uses the scientific features of your calculator to compute it for you. The second requires you to apply the formula yourself using the calculator only to do the actual computations.

TASK 13

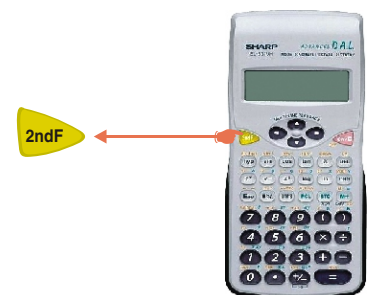
Calculating the standard deviation the easy way

Note if you have not actually deleted the data from calculating the mean it will still be in the calculator - in which case you can start from point 6 below.

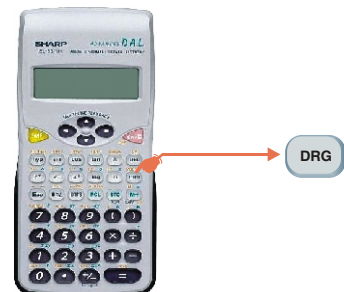
1. Switch the calculator on



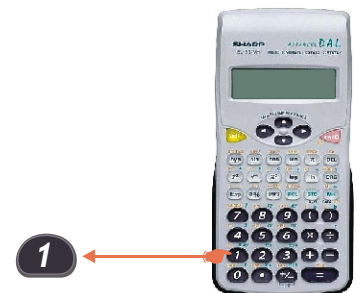
2. Press the key marked 2ndF (second function)



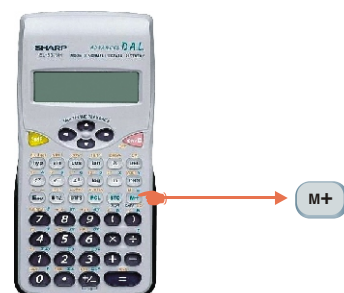
3. Press the key marked DRG



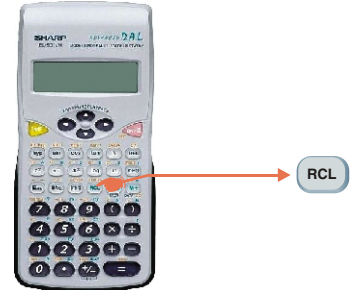
4. Press 1



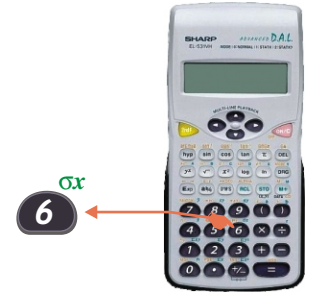
5. Enter each of the scores in the set of data you used for the previous task, each one followed by pressing the DATA key marked M+ (DATA is in black underneath).



6. Press the key marked **RCL** (**ALPHA** above this is its second function). You will see **ALPHA** appear at the top of the screen.



7. Press the key marked **6** (**σx** above and to right in green).



Write down your answer



At this stage it is probably hard for you to know if the answer makes sense. However if the data is normally distributed then nearly all the scores will fall within about 2 standard deviations of the mean. So if you think of there being about 2 standard deviations either side of the mean this implies that there are the equivalent of about 4 standard deviations in the range of scores. Remember this is just an approximate process for checking if your answer makes sense; it is not an exact measure. So if the range of scores is approximately 32 (difference between 2 and 34, you would expect the standard deviation to be about a quarter or a fifth of this, i.e. about 7. Is your result close to this? If not you need to check your calculation. Don't expect it to be exactly 7 but it should be in single figures and in that region. If it is much larger or much smaller then you have probably made a mistake.

Use the Course Forum to share your answer with others. Be brave, it's good to talk!

The longer way of calculating standard deviation is very laborious because you have to work out how much each score differs from the mean and then square each of these scores and then add them together, before you do anything else with them. We will leave it up to you if you want to do this but would advise that it will enhance your understanding of the mathematical process. So if you want to do standard deviations the harder way on to the next task but remember it is optional!

TASK 14
(OPTIONAL)

Calculating the standard deviation the slightly harder way

1. **Make sure the calculator is in “ordinary” mode (2nd function → MODE ⇒ 0). Use the calculator by entering each score in turn and pressing the x^2 button and then the = button.**



2. **To complete the first column deduct each score from the mean (you have already worked out the mean)**
3. **To complete the second column square this new number (mean minus score, squared)**
4. **Now sum all the squares of the differences from the mean and put the total in the last box.**

Score	Mean minus score	(Mean minus score) Squared
34		
32		
27		
26		
22		
20		
19		
18		
18		
16		
16		
14		
14		
14		
14		
14		
14		
14		
13		
13		
13		
13		
12		
12		
12		
12		
12		

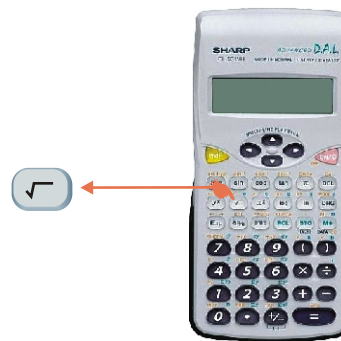
Score	Mean minus score	(Mean minus score) Squared
11		
11		
11		
11		
10		
10		
10		
8		
7		
7		
7		
6		
6		
6		
6		
6		
5		
5		
4		
4		
4		
4		
3		
3		
3		
2		
Total of squared differences from the mean		

5. **Divide this total by the number of scores. Enter your result here. Note this is the variance!**



Now to find the standard deviation you need to find the square root of this number. The easiest way to do this is to press the square root function button on your calculator and then enter the number followed by =. Finding square roots by hand is difficult! If you are old enough to remember, like us, it's something we used to use log tables for.

6. **Press the Square Root button marked $\sqrt{\quad}$**



7. **Then enter your number above**

Finally press = 

The answer should be exactly the same as when you calculated it earlier. If it isn't check them both against the answers file in the Resources Section (Unit 2) of the RealTraining website.

The meaning of standard deviations

So now you can calculate the standard deviation. But what does it mean? In effect it is simply a standard way of measuring the extent of the variance from the mean of a set of scores.

In practice this is useful because in a normal distribution (more of this in the next competency) the amount of people who score within one standard deviation from the mean is actually predictable. As it is for two standard deviations.

So if we know the mean and standard deviation of a normally distributed set of scores then we know what proportion of people would be in each of the segments, i.e. the proportion of cases that would occur between the mean and a particular deviation.

A normal curve with standard deviation units either side of the mean

This is perhaps best illustrated by an example. In a normal distribution approximately 34% of cases or scores will fall within one standard deviation (SD) of the mean. That is 34% one SD above the mean and 34% one SD below. Thus 68% (34×2) of all scores fall within one SD of the mean. The normal curve below and on page 27 illustrate this.

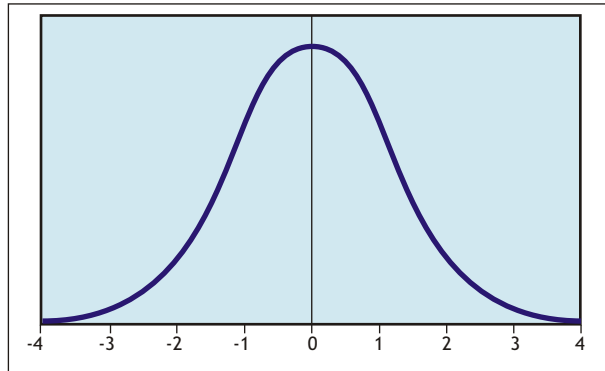


figure 2.3

The standard deviation is important because without it you do not know how well a person has done in relation to the rest of the population unless you can ascertain how far the score varies from the mean (average). You may know it is better than average but by how much? This is one of the most fundamental mistakes that is made in test result interpretation. The assumption is sometimes made that because a score is above the average it indicates much superior performance than the average when in fact it may be that it is well within one standard deviation from the mean and cannot be relied upon to indicate particularly superior performance.

Take the example of two tests both with a mean of 50. One has a standard deviation of 5 while the other has a standard deviation of 15. A score of 65 means completely different things on the two tests but it is easy to assume it is the same.

On the first ($SD = 5$) a score of 65 is 3 whole standard deviations above the mean and would only be achieved by less than 1% of a normal population - a very high performance.

On the second however ($SD = 15$) a score of 65 is only one standard deviation above the mean and would be achieved by approximately 16% of a normal population - good but not in the same league as the above example.

You will remember also from the histograms that you drew that the bulk of scores within a normal distribution fall around the average. In fact as you will see as we progress onto further examinations of the normal curve that bulk falls within one standard deviation of the mean and is generally considered to be the average range. There is some variation in test construction but by and large if the result falls within one standard deviation of the mean then one has to be very careful before drawing too many conclusions on that basis alone.

There will be much more on normal distributions later in this unit but it is time to consolidate your learning about standard deviations.

TASK 15

Using the book "A Psychometrics Primer" by Paul Kline which was provided with your course materials look up all the references to "variance" and "standard deviation", particularly in chapter 3 "The Characteristics of Good Psychometric Tests".

You could also try typing "define: standard deviation" into Google on the Internet and see what you get. Print one or two of them for reference.

Using all available information make some notes about standard deviation. You will need them for a paper you will write at the end of this unit.



Mean median and mode

We said earlier that there is more than one type of average. The one we have concentrated on so far is the arithmetic mean as this forms the basis for calculating the standard deviation. It is the mid point of a distribution as long as the distribution is normal which has been true for the scores on which you have been working so far. In fact psychometric tests are standardised in such a way that they do reflect the normal distribution so that the statistical operations that make them useful can be done.

It is possible to have distributions that are skewed to one side or the other. For you to have full understanding of the normal distribution it is important to consider these skewed types of distribution and the effect they have on the averages that can be computed.

There are three types of average: the mean, the median and the mode.

- Mean The mean is the sum of scores divided by the number of scores.
- Median The median is the score that divides the top 50% from the bottom 50%, i.e. half the scores are above and half are below but it takes no account of how far they deviate from the average.
- Mode The mode is the most frequently obtained score, i.e. the one most often recorded.

You can find further explanations of Mean, Median and Standard Deviation at <http://www.robertniles.com/stats/>.

TASK 16

In our data set that you used for TASKS 10-15 you have already calculated the arithmetic mean.

Now you should identify the median and the mode.

Remember the median will be the score that, if you line all the scores up in order of size, half will be to the right and half to the left. It doesn't matter how big or little they are in comparison to the median, just as long as there are half below and half above.

The mode is simply the value most often seen. It is likely to be in the middle somewhere so it is usually easy to find unless the distribution is very skewed indeed in which case you might find it towards one end.

Record them below

Normal curve

When the scores are distributed normally then the shape of the distribution is symmetrical; it looks roughly the same above the middle as it does below it. In this type of distribution, a unimodal distribution, i.e. "normal", the mean, the median and the mode have roughly equal values.

Positive skew

This is illustrated by the graph below

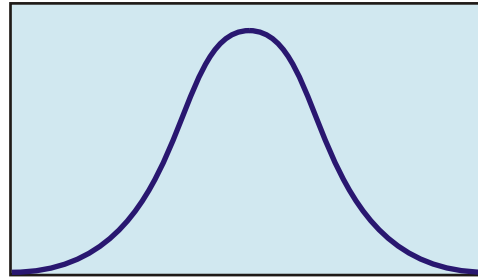


figure 2.4

When the scores are distributed in such away that more people obtained scores at the lower end of the range then the distribution is positively skewed because the mean is shifted up the scale because the scores above the mode are further from it and therefore have a disproportionate impact on the mean. In plain terms this means that scores above the mode (i.e. the most common score) are bigger in a positive direction than the scores below the mode in a negative direction. This is why it is called positively skewed even though the bulge appears to be towards the low scores.

This is illustrated by the graph below

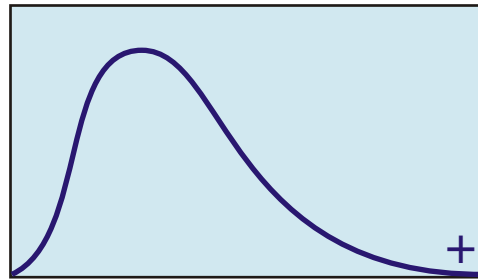


figure 2.5

Negative skew

When the opposite is true, i.e. the scores are distributed in such away that more people obtained scores at the upper end of the range then the distribution is negatively skewed because the mean is shifted down the scale because the scores below the mode are further from it and therefore have a disproportionate impact on the mean. In plain terms this means that scores below the mode (i.e. the most common score) are bigger in a negative direction than the scores below the mode in a positive direction. This is why it is called negatively skewed even though the bulge appears to be towards the high scores.

This is illustrated by the graph below

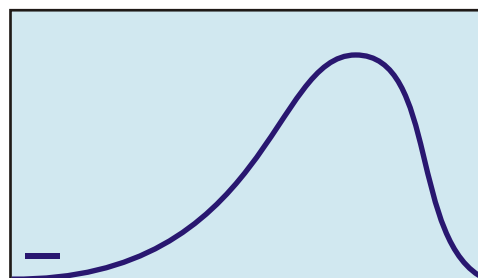


figure 2.6

TASK 17

In some ways this seems counter-intuitive so to convince yourself, quickly calculate the mean, median and mode for the following set of numbers.

1, 1, 2, 2, 2, 3, 3, 4, 5, 7, 8

What type of distribution is this?

Now do the same for this set of numbers.

1, 2, 3, 4, 5, 5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10

What type of distribution is this?

You might want to share your results on the Course Forum. Get some reinforcement for your correct answers!

Answers can also be found in the resources section of the RealTraining website.

Further reading

Statistics without Tears by Derek Rowntree

- Chapters 3 & 4. A simple good value for money paperback. Costs around £9.

Statistics for Dummies by Deborah Rumsey

- Chapter 5. Better presented than *Statistics without Tears* and probably easier to understand. Costs around £14.

Statistics for the Utterly Confused by Lloyd Jaisingh

- Chapter 2. Lots of diagrams, little text. Costs around £10.

Competencies covered in this section of unit 2

Section 2

Sample statistics

- 2.4 Describe the relationship between the Standard Error of the mean of a sample of test scores and the size of the sample.
- 2.5 Explain how the variation amongst a standardisation sample of test scores and their mean can be used to determine the level of confidence that we can have in people's scores on that test.

In other words by the end of this section you should be able to demonstrate that the standard error of the mean decreases as sample size increases and to describe what confidence limits are. You should be able to calculate by whatever means (e.g. calculator, computer program etc.) the standard error of the mean and confidence limits for a sample of means, given the mean, standard deviation and sample size; and the 68% and 95% confidence limits.

Standard Error of the Mean

Note that this competency could be considered one of the most challenging competencies in the whole course. We could have left it to the end but instead we have included it here in the order chosen by the British Psychological Society. If this is not clear on first reading we suggest you either come back to it later in the course or use the course forum on the website. This is the kind of information that we rarely use on a day to day basis but it is worth understanding at least once in our lives!

Imagine that the scores you have obtained for a group of people is not the only group of scores you could have obtained. The test could have been administered to any group of similar people. Theoretically the number of groups is infinite although in practice there are only so many groups that you could test with similar relevant characteristics, e.g. age, year group etc.

The point is that the group you did test is not unique and a similar group of people could get a slightly different set of results. Would the shape of the graph and consequently the mean and standard deviation be the same or would there be some variation? The answer is assumed to be that some variation would occur. This is called sampling error and reflects the fluctuation that would occur due to the fact that you have actually administered your test to a sample of possible people not the whole population.

So how much is the error? To establish this statisticians make use of the concept of the sampling distributions of the mean. In other words how would the means of each comparable sample vary?

Fortunately this variation is subject to the same statistical rules as the measurements of the scores in the first place. As long as the sample size exceeds about 30 the variation in the sample means is a normal distribution with a mean and a standard deviation. This standard deviation of the means is called the **standard error of the mean**.

The standard error of the mean is calculated simply by dividing the standard deviation of the population by the square root of the number in the sample; clearly the bigger the sample the smaller the error. This standard error of the mean can then be used to identify how confident one is that it represents an accurate score. This will be covered in the next competency.

TASK 18

On a reading test you obtained a mean reading age of 8:6 years. (You will see this written in several ways: 8 years 6 months or 8y 6m for example.) The standard deviation of the test was 10 months. 132 pupils took the test.

Calculate the standard error of the mean by dividing the standard deviation by the square root of the sample size.

Even if you had not used calculators before you should now be familiar enough with them to use the square root key marked ($\sqrt{\quad}$).

Check that your answer makes sense. Would you reasonably expect the standard error of the mean to be greater than the standard deviation? It could only be that if the square root of the sample was actually less than 1. And that could only be the case if the sample size itself is less than 1 - which is clearly impossible. So the implication is that the standard error of the mean must be less than 10 months. In practice it should be substantially less than 10 months.

If you think about it, even if you do not know the square root of 132 you will probably know the square root of a number near it; 100 for example or maybe 121.

Let's take 100 as an example. The square root of 100 is 10. So if the sample size had been 100 the standard error of the mean would be calculated by:

$$\begin{aligned} Se_{\text{mean}} &= \frac{SD}{\sqrt{N}} \\ &= \frac{10}{\sqrt{100}} \\ &= \frac{10}{10} \\ &= 1 \end{aligned}$$

Note: SD is another way of writing standard deviation.

So in this rough worked example the standard error of the mean is 1. That means that the standard error in the measurement is 1 month plus or minus the score obtained.

With a sample size of 132 you would expect the outcome to be about the same - slightly higher or slightly lower? Again with apologies to mathematicians out there who are feeling vaguely patronised by now "the higher the sample size the lower the error". So your answer above should be lower than 1. This is not only intuitively correct but also mathematically consistent since the higher the number you divide by the lower the result.

We hope this by now makes sense even to those who are terrified of maths and thought they had left it behind long ago. If you are still having difficulties you can always ask a colleague to go through it with you or post questions on the Course Forum of the Real Training website.

Confidence Intervals

The work you have just done on standard error of the mean has a direct bearing on the confidence with which you can state the score of any individual.

Because as we said in the previous competency the standard error of the mean is also a normally distributed range it obeys the same rules of distribution (see competency 2.6 for an explanation of the normal distribution). In other words one standard error from the mean accounts for 34% of the variance either way. That is 68% of scores would occur within one standard error; 96% of scores would occur within two standard errors of the mean.

In fact the most commonly used confidence level is the 95% level. This is the point at which the scores would occur within 1.96 standard errors of the mean. For all practical purposes this is close enough to 2 standard errors though in test construction the level of accuracy in these computations is precise.

In our example above:

A reading test with a mean of 8:6y

A standard deviation of 10months

A standard error of just under 1 month (you should have got an answer of about 0.87 but this can be rounded up to 1)

We can now see that for 95% confidence we can quote any score within a range of 2 standard errors either side of the actual score. To do this of course you need to know what the standard error is.

So for an obtained score of, let's say 7:8y, you could quote a range of 7:6y to 7:10y with 95% confidence.

Generally statistics would not allow you to quote anything less than a 95% confidence, but what does that actually mean? Basically it means that you would only expect to be wrong once in twenty times. This level of error is considered to be acceptable.

Of course you could be even more accurate by extending the confidence interval to three standard errors. In these circumstances you would then be over 99% confident that the real score fell within the range you have quoted but at this point in some tests the range can be come too great for practical use.

The use of the 95% confidence level is then to some extent a compromise. It provides an acceptable level of accuracy set against a reasonable level of certainty.

We hope you will never see a simple score quoted on a test result without asking what the confidence level is and thereby perhaps challenging the assumptions that are sometimes made on behalf of single scores.

TASK 19

Complete the following table to identify the relevant range of confidence for scores on the test we have been using above.

Just to remind you:

A reading test with a mean of 8:6y

A standard deviation of 10months

A standard error of just under 1 month (you should have got an answer of about 0.87 but this can be rounded up to 1)

Obtained score	68% confidence level	95% confidence level	99+% confidence level
7.2y			
9y 11m			
6.9y			
8.0y			
9y 2m			
8y 7m			

Standard Error of Measurement

The data we have used for the last two competencies actually provides for a fairly small standard error. There is also another method for calculating standard error and this is called standard error of measurement or SE_m . This is the method more typically used in test construction.

This method is based on the test - retest reliability of the test. You will learn more about this concept later but suffice to say at the moment it is about whether a test's results are the same if the test is used again with the same subject - test and retest. In other words does the same person get the same result if they take the test again? Clearly one has to allow for having "learnt" the test but in broad terms a test should obtain similar results each time it is used. If it didn't would you have any confidence in the results? If a pupil scored 40 out of 50 one time and then 20 out of 50 the next what would this mean? Is the test so boring he "switched off"? Is there a high chance factor and lots of guessing in the test? He had forgotten? Or maybe the test isn't measuring what it says it is measuring. Anyway without going into too much detail about test-retest reliability, as this will be covered later it is pertinent to raise here the way in which standard error can be measured.

In effect standard error of measurement is another way of indicating the confidence with which you can quote scores obtained on a test. What you have learnt about confidence intervals above is unchanged but the way of measuring the standard error is different. Both are used by statisticians and for your purposes it is not particularly important which method has been used as long as you can identify what the standard error is and what this means for quoting the confidence level of your obtained score. Effectively the standard error, however it is measured, sets the confidence limits for an obtained score.

So let us examine another example, one with a bigger standard error.

Imagine that two pupils obtained scores of 93 and 109 on a test with a mean of 100. At first sight it looks like the second pupil has outscored the first by a considerable margin and it would be easy to assume this to be the case.

However now add in the information that the standard deviation is 15 and the standard error is 5.

First of all both these scores fall within one standard deviation of the mean and as such are within the central band of the 68% of the population that do so.

Secondly if you apply the 95% confidence level then you have to apply a plus or minus 10 to the scores obtained to be 95% confident that the score lies within that band.

Then the score of 93 falls within the range 83-103 and the 109 falls in the range 99-119. Clearly these overlap and you cannot be sufficiently statistically certain that these two pupils are actually different in respect of their scores on this test. It seems counter-intuitive to see scores like this and not say they are different but that is why tests have to be used with such care.

As it happens a standard error of 5 is high for a well researched test. More realistically a standard error of 3 might be obtained with a good test and as test-retest reliability gets better the standard error will fall.

In the example above with a standard error of 3 the scores of 93 and 109 have ranges of 87-99 and 103-115. In this case their ranges do not overlap even accounting for the two standard errors either side. You can then be 95% confident they are different and you can begin to consider why.

However had they been 94 and 105 you could not be so certain as their ranges do still overlap (88-100 and 99-111).

You can see why the standard error is so important when drawing conclusions about test results.

TASK 20

The test scores below have been obtained on a test with a mean of 100, a standard deviation of 15 and a standard error of 3.

Complete the table filling the range of confidence at the 95% level and identify whether the scores in columns 1 and 2 are actually different with 95% confidence.

Score 1	Score 2	Range of score 1 (95% confidence level)	Range of score 2 (95% confidence level)	Can the scores be considered to be different?
88	95			
115	126			
98	108			
90	104			

In reality you look up the confidence intervals using tables in the back of a test manual.

Competencies covered in this section of unit 2

Section 3

Normal curves

- 2.6 Understand of the properties of the Normal Distribution and their relevance to measurement in general.
- 2.7 Use a set of statistical tables to establish the percentage of cases likely to fall below a particular test score.
- 2.8 Demonstrate how raw test scores can be converted into any of the scales frequently used to measure test performance (e.g. percentile scores, z-scores, standard scores, T-scores etc.).

The Normal Distribution

The normal distribution or normal curve is a mathematical concept that depicts a hypothetical, bell-shaped, symmetrical distribution of scores. It has not arisen by magic and it does not have magical properties. In fact it needs some demystification.

The normal curve merely represents the distribution to be expected when large numbers of random factors can influence a score. It is, in itself, derived from probability theory.

Perfectly symmetrical distributions rarely occur in real life but the greater the number of participants the more likely the phenomenon is to be seen. As it happens many distributions of human performance, attributes and characteristics closely resemble the normal distribution and so its mathematical properties are used to construct test norms.

It is this way round rather than the other that is the important thing to remember. Human performance happens to fit the profile of the normal curve rather than tests have been constructed to make this shape for the sake of convenience. Whilst the existence of a perfect curve is rare in nature the more trials you do the closer you get to the shape of the curve.

So well constructed tests use large samples, as their standardisation sample, to make sure they actually do reflect as far as possible the population's normal distribution.

You will note that normal distribution, normal curve and bell-curve are often used interchangeably. That is ok and you should know each term, as you will come across them all.

It is possible to illustrate the phenomenon of the greater number of trials getting closer to the expected distribution in a number of ways.

If you want to do something that resembles the normal curve then measure the heights of everyone you know and plot them on a histogram. We are not suggesting you actually do this but imagine how it would look. There will be a few at the extremities and many in the centre. But even if you do this for 30 people there will be slight anomalies. You may happen to know a lot of tall people and at the top end of the graph there will be a bump that should not be there. But if you increase the sample size and include people you do not know then the bumps will eventually be smoothed out. If you measured a thousand people then you will be getting pretty close to the normal curve when you plot your frequency distribution (e.g. histogram).

Medical authorities, which need to judge growth in young children, have access to charts and graphs based on thousands of people so that the distribution of height is so close to the normal curve that you would not know the difference.

To illustrate the phenomenon of how sample size eventually leads to a consistent distribution you could try the following simple experiment.

Take a coin and toss it four times. Record how many heads and tails. You might get two of each but there is a fair chance that you would get three of one and one of the other - or even four of one and none of the other.

But if you toss it 50 times you will get closer and closer to half and half, heads and tails. If you toss it 1000 times the split will be so close to 50% of each as to be infinitesimal. It might not be 500/500 but the error will be so small as a proportion of the total that the percentages will be within a whisker of 50%. If you toss it a million times ... you will get a sore wrist, but you get the point.

This of course is not a normal distribution but it illustrates the point that the bigger the sample the closer you get to the probable distribution. Because that is true of normal distributions as well it informs the way that tests are created.

TASK 21 *Get some definitions of normal distribution, normal curve or bell-curve by typing them into Google or by looking them up in the indexes of books on the reading list. Remember to type "define:" into Google and then the term you are researching. Alternatively you can also type "normal curve" as the main search term.*

Make some notes on the definitions.

TASK 22 *Search the Internet for pictures of normal curves. One way of doing this is to use www.google.co.uk and use the images tab. Make a list of the human characteristics that psychologists and educationalists believe are normally distributed. Post your list on the Course Forum. Read and discuss other people's lists. You may also want to post the link to a particularly good website about the normal distribution. There are lots of them.*

Areas under the normal distribution curve, using statistical tables and the transformation of scores

Raw Scores

Let us start by defining the various types of scores.

The raw score is the score you give the person taking the test. Raw scores in education are generally interval scales of measurement (see competency 2.1). They have rank order, equal intervals (often just 1) but no absolute zero. They are of limited use because they do not tell you how good or bad a result actually is. Without standardisation it is impossible to know whether 35 out of 50 is good, bad or average. This is OK for a criterion-referenced test (see competency 1.8) but not much use if you want to make comparisons with the rest of the population.

Standardised scores

Standardised scores are used where it is important to determine an individual's performance on a test relative to the performance of others who have taken the test or indeed could have taken the test. This clearly could be a large population. They also allow you to compare an individual's performance on one test with his or her performance on another test. Standard scores can be derived from raw scores using the properties of the presumed underlying frequency distribution of scores i.e. the normal distribution. Examples of standardised scores include T-scores, z-scores, stens and stanines. The most popular of these are described below. Those that aren't do not need to concern you for this course though you are of course encouraged to look them up if you wish and you will come across references to them in some standardised tests. These scales all have a predetermined mean and standard deviation. They can also be easily converted from one to another. All traditional psychometric tests are standardised and should have norm tables for different groups accompanying them.

Percentile scores

Percentile scores are ordinal rather than interval. As such they have rank order but they do not have equal intervals. In other words the percentiles if represented on a scale would not be in equal spaces; they would appear squashed in the middle and stretched at the ends. Percentiles can also be easily derived from standard scores and will also accompany many traditional norm-based psychometric tests. Percentiles cannot be used to calculate means, standard deviations etc. and make no assumptions about the normal distribution of the scores.

One of the most common transformations based on the normal distribution is the **z-score**.

The z-score is simply raw score minus the sample mean divided by the standard deviation.

$$z = \frac{(x - \text{mean})}{SD}$$

To illustrate this with an example let's use the same sort of data we had before. A score of 89 on a test with mean of 100 and standard deviation of 15.

$$\begin{aligned} z &= \frac{(89 - 100)}{15} \\ &= \frac{-11}{15} \\ &= -0.73 \end{aligned}$$

In effect the z-score tells you how many standard deviations the score is away from the mean expressed as a decimal. You can deduce that a score of 85 would have given a z-score of 1, a score of 115 a z-score of +1, a score of 130 a z-score of +2 ... and so on.

It is useful to put in the + if it is a positive z-score to avoid any confusion. Clearly a z-score that is positive means a score better than the mean and you don't want to miss this point.

z-scores are useful primarily because they provide data in a standardised format that makes it comparable with any other data on the normal distribution and they can be easily converted into percentiles (or centiles as they are often called).

A variation of the z-score is the T-score, which has a predefined mean of 50 and a standard deviation of 10.

The T-score is calculated by multiplying the z-score by 10 and adding 50. This means that the results are always positive and are more likely to make common sense.

Continuing our example above ... a score of 89 on a test with mean of 100 and standard deviation of 15.

$$\begin{aligned} z &= \frac{(89-100)}{15} \\ &= \frac{-11}{15} \\ &= -0.73 \end{aligned}$$

$$\begin{aligned} T &= (z \times 10) + 50 \\ &= (-0.73 \times 10) + 50 \\ &= 42.7 \end{aligned}$$

TASK 23

Calculate the following z-scores and T-scores from the following raw scores. You can assume a mean of 100 and a standard deviation of 15.

Raw score	z-score	T-score
124		
104		
139		
86		
71		

Because the shape of the normal curve can be measured and the area under the curve can be obtained then for any standardised test the percentage of subjects performing at each level can be identified. This is how percentiles are established.

They are in simple terms an expression of the area under the graph to the left of the line at which the score occurred. If you had to work this out each time you would need to be competent in differential calculus. Luckily the pattern is predictable because tests are designed in such a way as to be clearly related to the normal distribution. So the percentiles can be seen approximately on graphs and looked up from tables for more precision.

The figure below shows you the relationship of the normal distribution to various types of standard scores.

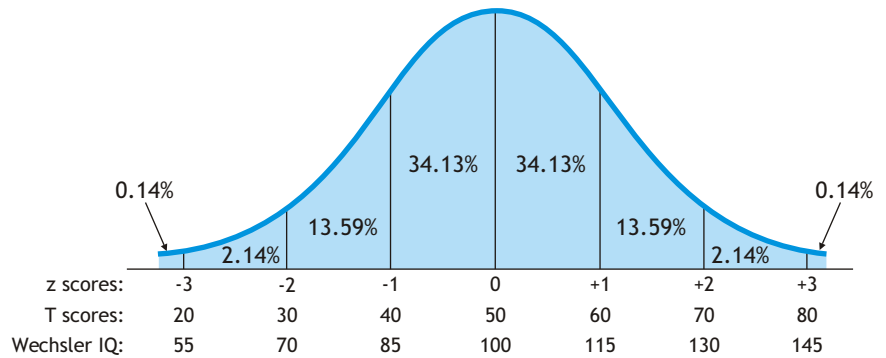


figure 2.7

Examples of standard scored in a normal distribution, with approximate percentages.

z-scores and Percentiles

The tables below allow you to convert z-scores into percentiles.

z-score	Percentile	z-score	Percentile	z-score	Percentile	z-score	Percentile
-4.0	0.01	-1.0	15.87	0.0	50.00	1.1	86.43
-3.5	0.02	-0.95	17.11	0.05	51.99	1.2	88.49
-3.0	0.13	-0.90	18.41	0.10	53.98	1.3	90.32
-2.9	0.19	-0.85	19.77	0.15	55.96	1.4	91.92
-2.8	0.26	-0.80	21.19	0.20	57.93	1.5	93.32
-2.7	0.35	-0.75	22.66	0.25	59.87	1.6	94.52
-2.6	0.47	-0.70	24.20	0.30	61.79	1.7	95.54
-2.5	0.62	-0.65	25.78	0.35	63.68	1.8	96.41
-2.4	0.82	-0.60	27.43	0.40	65.54	1.9	97.13
-2.3	1.07	-0.55	29.12	0.45	67.36	2.0	97.72
-2.2	1.39	-0.50	30.85	0.50	69.15	2.1	98.21
-2.1	1.79	-0.45	32.64	0.55	70.88	2.2	98.61
-2.0	2.28	-0.40	34.46	0.60	72.57	2.3	98.93
-1.9	2.87	-0.35	36.32	0.65	74.22	2.4	99.18
-1.8	3.59	-0.30	38.21	0.70	75.80	2.5	99.38
-1.7	4.46	-0.25	40.13	0.75	77.34	2.6	99.53
-1.6	5.18	-0.20	42.07	0.80	78.81	2.7	99.65
-1.5	6.68	-0.15	44.04	0.85	80.23	2.8	99.74
-1.4	8.08	-0.10	46.02	0.90	81.59	2.9	99.81
-1.3	9.68	-0.05	48.01	0.95	82.89	3.0	99.87
-1.2	11.51	0.0	50.00	1.0	84.13	3.5	99.98
-1.1	13.57					4.0	99.99

figure 2.8

TASK 24

Using the figures 2.7 and 2.8 compare various types of standard scores.

Complete the following table

Percentile	z-score	T-score
84		
	-1.4	
8		
	+2.1	
46		

Now that you have completed this unit you should be able to demonstrate an understanding of the relationship between percentile scores, z-scores and T-scores. You should be able to convert different raw scores provided at the assessment day into percentiles and z-scores. You should also be able to use tables familiar to you to convert percentiles to z-scores and vice versa. You should also be able to derive standardised scores and percentile scores from raw data and to convert raw data to non-normalised z-scores and stanines and vice versa. You should be able to use norm tables to find percentile equivalents of raw scores and then to obtain z-scores and T-scores from normal distributions.

We have provided some practice tasks below that will be similar to those that you will encounter at the Competence Day.

TASK 25

Using the scores below

72, 73, 74, 77, 80, 82, 83, 84, 86, 87, 90, 91, 92, 94, 94, 95, 96, 97, 98, 98, 99, 100, 100, 101, 101, 101, 103, 104, 105, 107, 108, 109, 110, 111, 113, 114, 115, 116, 119, 122, 125, 126, 129

- Construct a histogram of the scores. You could use a 10 point interval starting at 70-79...
- Do the data look normally distributed?
- Calculate the mean and standard deviation.
- Calculate the standard error of the mean.
- Identify the range of confidence for each of the raw scores in the table below.
- Identify the range of scores that fall within one standard deviation either side of the mean.
- Do any scores fall outside 2 standard deviations from the mean?
- Work out the z-score for the raw scores in the table below.
- For those raw scores work out the percentile scores and T-scores.

Raw	Range of confidence (2 Sds)	z-score	T-score	Centile
84				
104				
115				
129				

- What conclusions could you draw about the performances of those people who recorded the highlighted raw score? Take care to use all the information now at your disposal, including the standard error of the mean.**

You can check your results on the website in the Resources Section (Unit 2). If you need further work on this use the further reading suggestions or post a question on the Course Forum. You might also like to discuss your conclusions on the Course Forum with other course participants.

FURTHER READING

- *A Psychometrics Primer* by Paul Kline. In particular chapters 2 and 3.
- *Modern Psychometrics* by John Rust and Susan Golombok. In particular the sections on standardisation and normalisation in chapter 5.
- *Statistics for Dummies* by Deborah Rumsey, chapter 8 and 9.

Later in this course you will be applying the knowledge you have gained in this unit to real tests. While it is fresh in your mind you might therefore like to consider the next task.

TASK 26 (OPTIONAL)

Choose a test that you want to learn to use. Have a look through the manual and pay particular attention to the scoring instructions. Make up a hypothetical raw score. Also choose a hypothetical age for the child. Now use the tables in the manual to convert the raw score into a percentile and any other relevant scores (e.g. Standard Scores). Make a note of the confidence levels and convert any scores to ranges.

Ask a colleague to check your results.

You have now covered all the competencies in unit 2. In many respects this is probably the hardest unit, particularly if you are not mathematically minded. So you are well on your way to completing this course. We have one final task for you that hopefully will help to consolidate your learning into one written piece.

TASK 27

SUMMARY TASK FOR THIS UNIT

With all the information you now have at your disposal, including the notes you have made from reading texts and from the Internet prepare a 500 word paper on the following:

"How standardised tests are scored"

or

"The key mathematical ideas behind standardised testing"

Upload it to the Real Training website. Do not use the Course Forum on the Real Training website. Use the Work Submission Area. 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.

You will receive feedback on your paper from the course tutors then you can decide whether to publish on the website for feedback from other course members.

Assessment of Unit 2

Assessment of the competencies in unit 2 is by a few short exercises at a Competence Day. See the Real Training website for more details.

When you attend please bring along a test with which you are familiar. One exercise will involve you demonstrating your competence in using the tables in the test manual by looking up scores from a raw score.