

April 2008/14

Counting what is measured or measuring what counts?

League tables and their impact on higher education institutions in England

Report to HEFCE by the Centre for Higher Education
Research and Information (CHERI), Open University, and
Hobsons Research

Appendix B

Standard statistical concepts, methods and processes used in the compilation and analysis of league tables

League tables and their impact on higher education institutions in England

Appendix B: Standard statistical concepts, methods and processes used in the compilation and analysis of league tables

Contents

	Page number
1. Means and variability	2
2. Standardisation	4
3. Correlation and correlation coefficients	7
4. Regression and multiple regression	9
5. Factor analysis	9
References	14

Introduction to Appendix B

Appendix B provides a guide to the statistical concepts, methods and processes used in the compilation and analysis of league tables. It has been prepared by Professor John Richardson from The Open University and an Associate of CHERI. It is aimed at non-statisticians and seeks to illuminate the analyses carried out in Appendix C and summarised in the main report.

1. Means and variability

When carrying out procedures to manipulate distributions of scores or other measures, statisticians are mainly interested in three properties: the *middle* of a distribution; the *variability* of a distribution; and the *shape* of a distribution. With regard to shape, they are mainly interested in the extent to which a distribution follows a *normal curve*; this is a particular kind of bell-shaped distribution that is well understood mathematically. It frequently tends to arise as a result of naturally occurring processes; an example is the distribution of the heights of adult males in a specific population. It can also arise artificially as the result of particular scoring procedures: for example, intelligence tests are often deliberately scored to produce distributions of IQs that follow a normal curve (see Figure 2).

The middle of a distribution can be defined in several different ways. We often refer to these in everyday language as measures of the 'average' of a set of scores. Most commonly, this term refers to the quantity that statisticians describe as 'the arithmetic mean'. There are other kinds of mean, but if statisticians just refer to 'the mean', this is the one they have in mind. It is defined as the total of the set of scores in a distribution divided by the number of scores in the distribution. For example, suppose that the scores obtained were 1, 2, 3, 4 and 5 (see Table 1). The total of these scores is 15, the number of scores is 5, and so the mean is $15 \div 5 = 3.00$.

Unless all the scores in a distribution are the same, most of them will be different from the mean. One way of characterising this variability is to measure the difference between each of the original scores and their mean. The *deviation* of a score about the mean is obtained by subtracting the mean from the original score. Scores greater than the mean will produce positive deviations, and scores less than the mean will produce negative deviations. For the example in Table 1, the deviations (shown in the second column) are obtained by subtracting the mean of 3.00 from each of the original scores.

Table 1: A worked example

	Original value	Deviation	Absolute deviation	Squared deviation	Standard score
	1	-2	2	+4	-1.41
	2	-1	1	+1	-0.71
	3	0	0	0	0.00
	4	+1	1	+1	+0.71
	5	+2	2	+4	+1.41
Total	15	0	6	10	0.00
Mean	3.00	0.00	1.20	2.00	0.00

One might try to measure the overall variability in a set of scores by taking the mean of all the deviations. However, the negative deviations will exactly counterbalance the positive deviations, so that the total of all the deviations will be zero, and the mean of the deviations (the total of all the deviations divided by the number of deviations) will also be zero. It can be shown that this will always happen, regardless of the scores one starts with. But it would clearly be nonsense to conclude that there was no variability for the example in Table 1.

The solution is to prevent the negative deviations counterbalancing the positive ones. In principle, this could be achieved in several different ways. For instance, one could just take the *absolute values* of the original scores (i.e. ignoring their signs). For the example in Table 1, the absolute values of the deviations are shown in the third column. In this example, the total of the absolute deviations is 6, and the mean of the absolute deviations is $6 \div 5 = 1.20$. However, it is hard to analyse the mathematical properties of the mean absolute deviation, and so statisticians take another approach.

This is to take the *square* of the deviations. This exploits the fact that multiplying a number by a negative number changes its sign. So a positive number multiplied by a negative number yields a *negative* result: for instance, $(+2) \times (-3) = (-6)$. However, a negative number multiplied by a negative number yields a *positive* result: for instance, $(-2) \times (-3) = (+6)$. In particular, the square of a negative number (that is, multiplying it by itself) yields a positive result: for instance, $(-2) \times (-2) = (+4)$. But the square of a positive number also yields a positive result: for instance, $(+2) \times (+2) = (+4)$. So the squared deviation will always be a positive number (or zero, if the deviation is zero). Consequently, the total of the squared deviations will be a positive number, and the mean of the squared deviations will be a positive number. For the example in Table 1, the squared deviations are shown in the fourth column. In this example, the total of the squared deviations is 10, and the mean of the squared deviations is $10 \div 5 = 2.00$.

The bigger the difference between the original score and the mean, the bigger will be the deviation in absolute terms (i.e. ignoring whether it is positive or negative), and the bigger will be the squared deviation. For the example in Table 1, the scores of 1 and 5 are farther away from the mean than the scores of 2 and 4, and so they yield larger squared deviations. The mean of the squared deviations therefore measures the overall variability of the distribution and is known as its *variance*. So the variance of the scores shown in Table 1 is 2.00.

This is a useful measure, but it has the disadvantage that it is not based on the same measurement scale as the original scores. Indeed, it may not be based on a sensible measurement scale at all. For instance, if the scores in Table 1 represent weights in kilograms, then the deviations represent the difference between each weight and the mean weight, again in kilograms. However, the variance is the mean of the squared deviations and is thus measured in 'squared kilograms'.

The obvious solution is to take the square root of the variance in order to get back to the original scale of measurement. (The square root of a number is the value which, when

multiplied by itself, yields the original number. For instance, the square root of 16 is 4, because $4 \times 4 = 16$.) The square root of the variance is known as the *standard deviation*. For the example in Table 1, the standard deviation is the square root of 2.00, which (to two decimal places) is 1.41 (because $1.41 \times 1.41 = 2.00$, ignoring rounding error). If the original scores are measured in kilograms, the variance will be measured in 'squared kilograms', but the standard deviation will just be measured in kilograms.

In other words, the standard deviation is a measure of the variability in a distribution of scores that is expressed in the same scale of measurement as those scores. Moreover, the mathematical properties of the standard deviation and the variance have been well analysed, which is why statisticians prefer to use them as measures of the variability of a set of scores about their mean. In many distributions, nearly all the scores will fall within three standard deviations either side of the mean, and most will fall within two standard deviations either side of the mean.

Two other properties of distributions are their *skewness* and their *kurtosis*. Skewness is a measure of the extent to which the distribution is symmetrical or asymmetrical. A positive value of skewness means that there is a relative preponderance of high scores (i.e. the distribution is positively skewed); a zero value means that the distribution is symmetrical about its mean; and a negative value means that there is a preponderance of low scores (i.e. the distribution is negatively skewed). Kurtosis is a measure of the extent to which the distribution is relatively peaked or flat in comparison with a normal curve. A positive value means that the distribution is relatively peaked in comparison with a normal curve, and a negative value means that the distribution is relatively flat in comparison with a normal curve. Distributions that follow a normal curve have zero skewness (because they are symmetrical) and zero kurtosis (just because they are normally shaped).

2. Standardisation

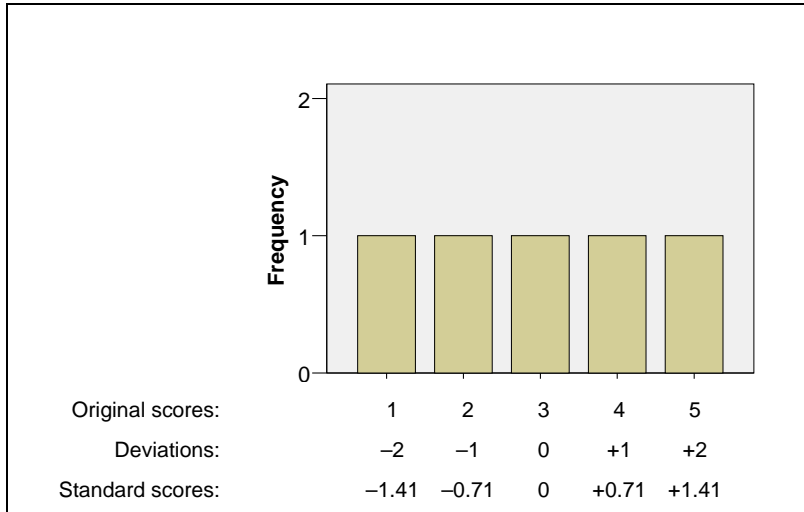
This term is used in a wide variety of ways. In general, it means to express different distributions of scores in a standard form. The most common form of standardisation used by statisticians is to transform a distribution of scores so that they have a mean of zero and a standard deviation of one.

The first step involves subtracting the mean of the distribution from each of the scores – in other words, taking the deviations of the scores. As mentioned above, the mean of the deviations is zero, so this is a transformation that changes the mean of the scores. However, the deviation of a deviation (i.e. the deviation about *its* mean) is the same as the deviation, since the mean of the deviations is just zero. So the standard deviation of the deviations is just the same as the standard deviation of the original scores. For the example in Table 1, the standard deviation of the deviations is just 1.41.

The distribution of deviations also has the same shape as the original distribution of scores. For the example in Table 1, each of the scores 1, 2, 3, 4 and 5 is obtained just once. Figure 1 shows a bar chart of these data; it consists of a row of five bars, all just

one unit high. This looks like a rectangle (in this case, one unit high and five units wide), and so it is known as a rectangular distribution. The deviations consist of the values -2 , -1 , 0 , $+1$ and $+2$. Figure 1 shows that a bar chart of the deviations also looks like a rectangle and has the same shape as the distribution of the original scores.

Figure 1: Frequency distribution of the data in Table 1

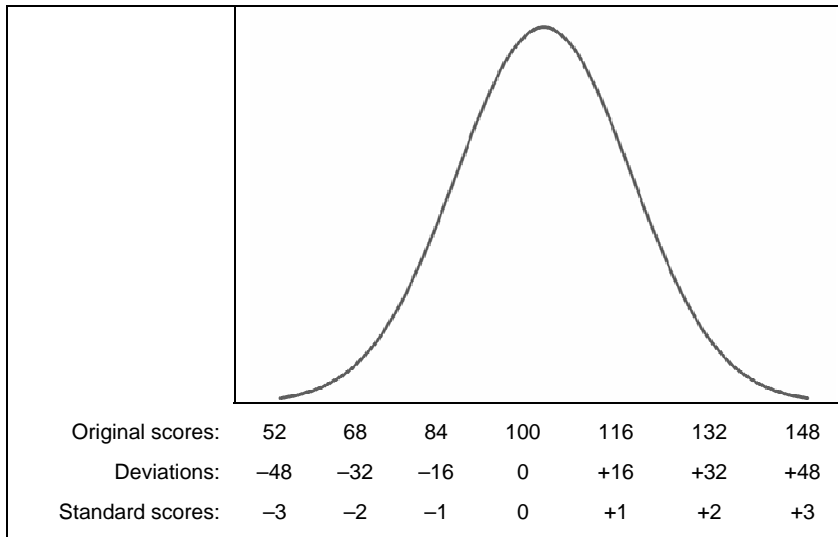


The next step involves dividing each of the deviations by the standard deviation. The transformed score obtained is known as a *standard score*. For the data in Table 1, the standard scores are -1.41 , -0.71 , 0.00 , $+0.71$ and $+1.41$, respectively. This transformation *does* change the standard deviation of the scores. In fact, the standard deviation of the standard scores is always one. So calculating standard scores turns any distribution of scores into one that has a mean of zero and a standard deviation of one. Going back to the example of a distribution of weights, the object that had a standard score of $+2$ would have a weight that was two standard deviations above the overall mean; the object that had a standard score of -1 would have a weight one standard deviation below the overall mean. However, taking standard scores does not change the shape of the distribution. For the example in Table 1, each of the standard scores -1.41 , -0.71 , 0 , $+0.71$ and $+1.41$ is obtained just once. Figure 1 shows that a bar chart of the standard scores also consists of a row of five bars, all just one unit high – in other words, a rectangular distribution.

As a second example, consider the distribution of IQ scores shown in Figure 2. It follows a normal curve with a mean of 100 and a standard deviation of 16. So a person with an IQ of 116 has an IQ that is one standard deviation above the mean. Once again, to standardise these data, the first step involves subtracting the mean of the distribution from each of the scores to obtain the deviations. The mean of the deviations is zero, so this transformation changes the mean of the scores. But the standard deviation of the deviations is still 16; the person whose IQ is one standard deviation above the mean (i.e. 116) has a deviation of $+16$, which is still 16 points above the mean deviation. So taking the deviations does not change the variability of the scores.

The next step involves dividing each of the deviations by the standard deviation (i.e. 16). A person with an IQ of 132 has a deviation of $(132 - 100) = 32$, and a standard score of $(132 - 100)/16 = 2$. This is just a way of saying that their IQ is two standard deviations above the mean. Similarly, a person with an IQ of 84 has a standard score of $(84 - 100)/16 = -1$. In other words, their IQ is one standard deviation below the mean.

Figure 2: An example of a normal curve



This transformation changes both the mean and the standard deviation of the scores. Once again, taking standard scores turns any distribution of scores into one that has a mean of zero and a standard deviation of one. However, Figure 2 shows that taking standard scores does not change the shape of the distribution. If (as in this example) the original distribution of scores is normally shaped, the distribution of standard scores will also be normally shaped. (In this situation, the standard scores are known as *z* scores, and their distribution is known as the standard normal distribution.) Conversely, if the original distribution of scores is *not* normally shaped, then the distribution of standard scores will not be normally shaped either.

The use of a mean of zero and a standard deviation of one is mainly for mathematical convenience, and researchers sometimes use other ways of standardising scores. For instance, clinical researchers often transform their data into *T* scores, which have a mean of 50 and a standard deviation of 10. In all these cases, working with standard scores enables the same individuals to be meaningfully compared on different measures, since they are transformed into a common scale. However, when statisticians talk about 'standardisation', standard scores are what they have in mind.

Our interviews with league-table compilers revealed some other usages of this term. In some cases, 'standardisation' refers to a process of mapping a set of scores onto a common scale (for instance, from 0 to 10, or from 0 to 100). The first step in this process is to identify the highest score and the lowest score and then to take the difference between these two scores, which is known as the *range*. For the example in Table 1, the highest score is 5, the lowest score is 1, and the range is $5 - 1 = 4$. If the intended scale

of scores is from 0 to 100, then the intended range is $100 - 0 = 100$. The next step is to take the difference between each score and the lowest score, divide by the actual range, and multiply by the intended range. Thus, the score of 3 maps onto $[(3 - 1) \div 4] \times 100 = 50$. The highest score maps onto $[(5 - 1) \div 4] \times 100 = 100$. The lowest score maps onto $[(1 - 1) \div 4] \times 100 = 0$. This transformation changes the mean and the standard deviation of the distribution of scores, but it does not change the shape of the distribution.

In other cases, league-table compilers have mapped sets of scores onto a scale with a maximum of 100, but where the minimum is greater than zero. This might have been achieved by expressing each original score as a percentage of the maximum score. For the example in Table 1, the maximum score is 5, and so the score of 3 maps onto $(3 \div 5) \times 100 = 60$. In this case, the highest score maps onto $(5 \div 5) \times 100 = 100$, but the lowest score maps onto $(1 \div 5) \times 100 = 20$, not zero.

There also appear to be cases where league-table compilers have 'standardised' a set of scores by replacing them with their ranks in an ordering from the lowest to the highest. If there are 100 scores on each variable, this will map them onto a common scale from 1 to 100. It will change the mean, the standard deviation, and the shape of the distribution of scores. (If all the scores are different, then the distribution becomes a rectangular distribution, since each rank will occur just once – like the scores shown in Table 1, which already represent a rank ordering.) This transformation assumes that the steps between successive pairs of ranks are equally important: for instance, that the difference between the highest score and the second highest score is as important as the difference between the second highest score and the third highest score.

League-table compilers have also used the term 'normalisation', which seems to be used in a number of different ways. In some cases, it refers to an adjustment to take into account variations in the size of institutions (for instance, citation counts may be divided by the number of staff). For statisticians, 'normalisation' most often refers to the process of transforming a distribution of scores into a normal distribution or even the standard normal distribution. As mentioned above, taking standard scores is not sufficient to achieve this: standard scores achieve a common scale in which the mean is zero and the standard deviation is one, but the shape of the distribution remains the same. In principle, more complex procedures could be used to transform the original scores or the ranks of those scores into a normal distribution. However, none of the variables used by any of the league tables we examined was normally distributed.

3. Correlation and correlation coefficients

Many statisticians are interested in studying the relationships between two or more sets of scores. A *perfect* relationship is one that can be represented exactly by a line, so that one score exactly predicts the other score, and vice versa. Most relationships in educational and social research are *imperfect* relationships, which can be more or less approximated by lines. A *positive* relationship is a direct relationship, where an increase in one variable tends to be associated with an increase in the other variable. (In a graph

of the relationship between the two variables, the line goes up from left to right.) A *negative* relationship is an inverse relationship, where an increase in one variable tends to be associated with a decrease in the other variable. (In a graph of the relationship between the two variables, the line goes down from left to right.)

The simplest relationships to study are *linear* relationships: that is, the pairs of scores more or less define a straight line when they are plotted as a graph (as opposed to *curvilinear* relationships, which are best fitted by more complex functions that appear as curved lines). From a theoretical point of view, the simplest assumption in practical situations is often that the relationship between two scores is a linear one, and there is usually no rationale for expecting the relationship to be more complex than this.

There are well-established procedures for finding the straight line that best fits such data, and this will be expressed as an algebraic equation relating the pairs of scores. For instance, one could arrive at the equation that best captures the linear relationship between height and weight in a group of people. This equation will depend on the exact form of the relationship, but also on how the scores have been measured. The equation representing the relationship between height in inches and weight in pounds will be different from the equation representing the relationship between height in centimetres and weight in kilograms. It is also not possible to ask directly whether a particular person occupies the same position in the distributions of height and weight.

One way to avoid this problem is to convert both measurements to standard scores. A person who is two standard deviations above the mean of the population in terms of their height will occupy the same position in the distribution of people's heights, regardless of whether their height is measured in inches or centimetres. Taking standard scores therefore provides a way of expressing the relationship between height and weight in terms of a common scale that is independent of how those variables were measured. This also provides a way to measure the strength of the relationship between the two variables.

A *correlation coefficient* measures the magnitude and the direction of the relationship between two sets of scores obtained from the same individuals. Correlation coefficients vary from +1, reflecting a perfect positive relationship, to 0, reflecting no relationship, to -1, reflecting a perfect negative relationship. Imperfect positive relationships produce correlation coefficients between 0 and +1; imperfect negative relationships produce correlation coefficients between 0 and -1.

The most commonly used correlation coefficient is the linear correlation coefficient Pearson *r*. It is sometimes called the product-moment correlation coefficient because of the way it is calculated. It expresses the extent to which individuals occupy the same relative positions in two different distributions of scores (for instance, whether someone who is two standard deviations above the mean in terms of their height is also two standard deviations above the mean in terms of their weight). Negative values of Pearson *r* mean that the person occupies the *opposite* positions in two different distributions. (For instance, someone who is two standard deviations *above* the mean in terms of their

weight might be two standard deviations *below* the mean in terms of how high they can jump.) There are other correlation coefficients, but when statisticians talk about ' r ' or just 'the correlation coefficient', it can be assumed that this is what they have in mind.

Pearson r is equal to the slope of the straight line that best fits the relation between two variables when they are expressed in terms of standard scores. So the variables themselves do not need to be converted to standard scores before determining the value of Pearson r . In other words, standardisation is unnecessary when carrying out analyses based on correlation coefficients. There is yet another way of interpreting Pearson r , which is that the square of r (r^2) is equal to the proportion of variation in one of the variables that is predicted or explained on the basis of its relationship with the other variable.

4. Regression and multiple regression

(Linear) regression is the use of linear relationships between two or more variables to predict the values of one of the variables on the basis of the values of the other variables. If there are just two variables, X and Y (e.g. height and weight), the linear relationship between them can be represented by the equation $Y = a + bX$, where Y is the predicted value of Y for a specific value of X , and a and b are chosen to minimise the errors in predicting Y . b is the *slope* of the line; a is the *intercept*, the point where the line crosses the Y axis in a graph. If the line passes through the origin, then $a = 0$ and the line simply takes the form $Y = bX$. If there is a perfect relationship between X and Y , the predicted values of Y will equal the actual values, there will be no errors of prediction, and r^2 will be equal to 1.

In multiple regression, there are two or more predictor variables, X_1 , X_2 , etc, and so the linear relationship with the predicted variable takes the form $Y = a + b_1X_1 + b_2X_2 + \dots$. The strength of the relationship between the predictor variables and the predicted variable is measured by the multiple correlation coefficient, R , and the square of the multiple correlation coefficient, R^2 , is equal to the proportion of variation in the predicted variable that is predicted or explained on the basis of its relationship with the other variables. All of the league table compilers indicate that they have defined an overall score for each institution as the weighted sum of the various indicators they have used. It follows that the relationship between the overall scores and the scores on the various indicators should take the form $Y = b_1X_1 + b_2X_2 + \dots$. Moreover, this should be a perfect relationship, so that $Y = Y$ and $R^2 = 1$. We used multiple regression to try to reconstruct the overall score for each institution on the basis of the published data.

5. Factor analysis

The purpose of factor analysis is to look for patterns in the correlation coefficients among a set of variables. If several variables are all very highly correlated with one another, then it is reasonable to assume that they are all tapping the same underlying construct. For

example, Richardson (1978) asked 160 students which hand they used, or preferred to use, for different activities: writing, throwing, cutting with scissors, playing with a racket or bat, brushing their teeth and striking a match. The correlation coefficients among their responses are shown in Table 2. They are all positive and high, indicating that most people – although not all – reported using the same hand for most of these activities. This suggests that it is sensible to talk about a single underlying dimension that might be called ‘handedness’.

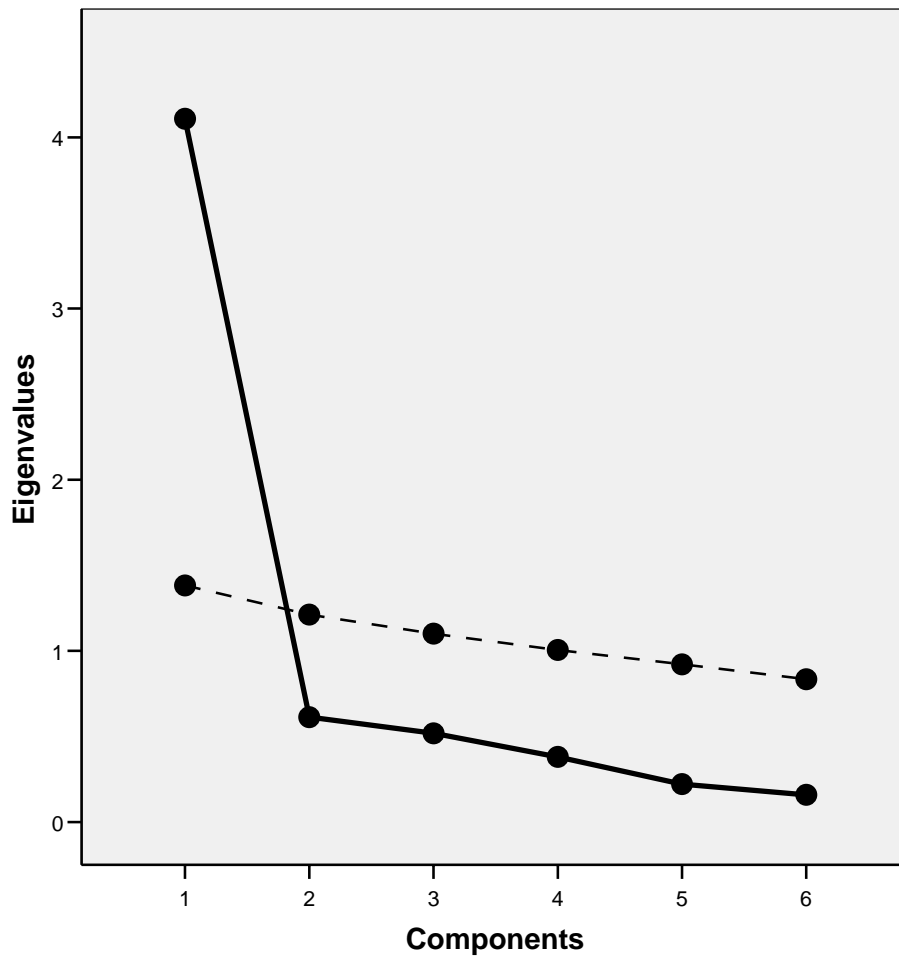
Table 2: Correlation coefficients among hand preference in six activities

Activity	1	2	3	4	5	6
1. Writing	1.00	.82	.66	.53	.71	.76
2. Throwing	.82	1.00	.61	.56	.61	.79
3. Cutting with scissors	.66	.61	1.00	.41	.60	.58
4. Playing with a racket or bat	.53	.56	.41	1.00	.53	.51
5. Brushing one’s teeth	.71	.61	.60	.53	1.00	.56
6. Striking a match	.76	.79	.58	.51	.56	1.00

The present project followed previous researchers such as Yorke (1997, 1998) in employing a factor-analytic procedure known as principal components analysis. This proceeds by finding a hypothetical trait (the principal component) that provides the best fit to all the observed variables. The ‘fit’ is defined in terms of the linear relationship between the trait and the standard scores of the original variables. The first principal component explains the greatest amount of variance in the standard scores. Each subsequent principal component reflects the best fit to the standard scores when previous principal components have been taken into account. The successive amounts of explained variance are known as the *eigenvalues* of the principal components. The solid line in Figure 3 shows the eigenvalues of the principal components for the handedness data shown in Table 2. For instance, the first principal component explains 4.109 of the six units of variation from the six standard scores; in other words, it explains $(4.109 \div 6)$ or 68.5% of the variation in the data.

This graph has a characteristic form and is known as a ‘scree plot’. The first part of the curve falls steeply, like a cliff, and represents genuine traits in the data. The second part is flatter, like the scree (or rubble) at the bottom of a cliff and represents random variation in the data. According to this visual scree test, there is just one hypothetical trait underlying the data in Table 2. However, the scree test is inherently subjective, and so researchers have looked for more objective criteria. One such criterion is to identify the number of principal components whose eigenvalues are greater than one. This is the default criterion used by many statistical packages, and it was the criterion used by Yorke (1997, 1998). For the data in Figure 3, the eigenvalues-greater-than-one rule also suggests that there is just one underlying trait.

Figure 3: Scree plot for the handedness data in Table 2



The rationale for this criterion is that if there were no relationships among the variables (i.e. all of the correlation coefficients were zero), then each variable would need to be represented by a separate component that only explained the variation in the standard scores of that variable, and so all of the eigenvalues would be equal to one. Nevertheless, if the data are merely a *sample* from a population in which there are no genuine relationships among the variables, the correlation coefficients will differ from zero purely as the result of chance variation, and the eigenvalues of the first few principal components will then be greater than one. As a result, the eigenvalues-greater-than-one rule in practice tends to overestimate the true number of components in a dataset (Cliff, 1988).

In fact, one can predict what the scree plot would look like if the data really were a sample from a population in which there were no genuine relationships among the variables. This technique is known as *parallel analysis*. The dotted line in Figure 3 shows the results of a parallel analysis of 1,000 random correlation matrices using a program written by O'Connor (2000). For the reasons just explained, the eigenvalues of the first few principal components are actually greater than one. The first principal component is the only one with an obtained eigenvalue which is greater than what would be expected

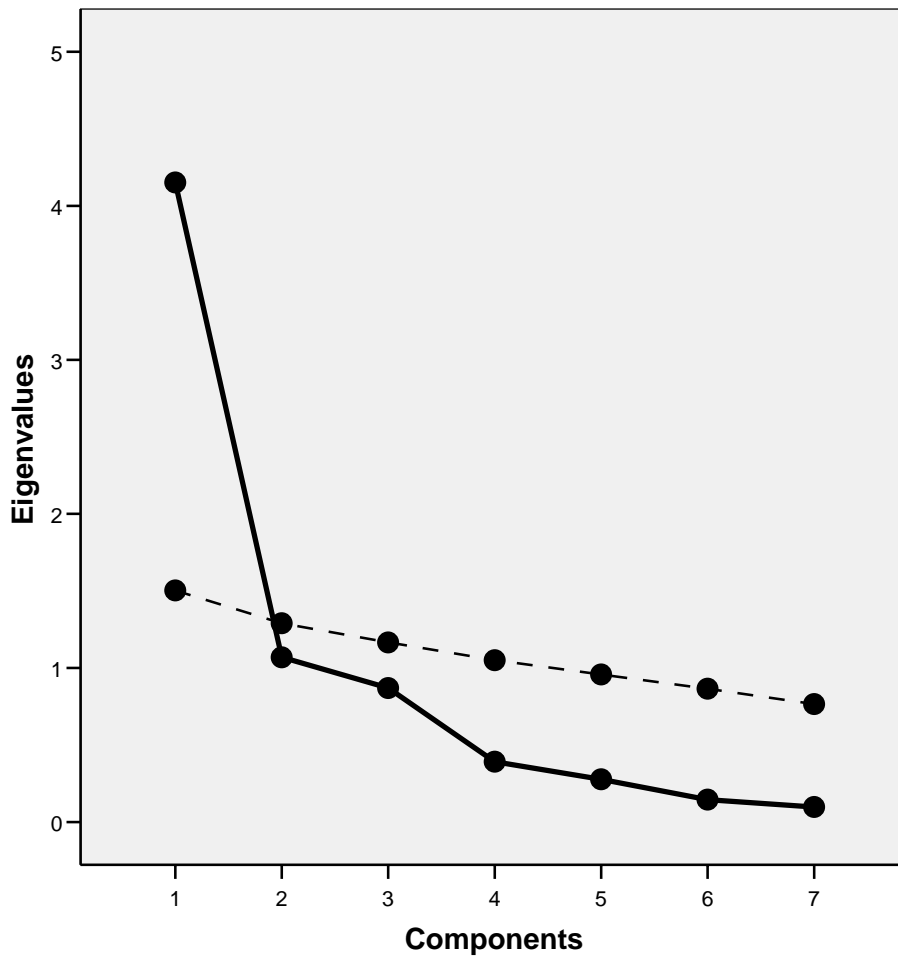
from purely random data, and this implies once again that just one component should be extracted from the dataset.

In this example, the scree test, the eigenvalues-greater-than-one rule and parallel analysis all yielded the same answer to the question of how many components should be extracted from a dataset. Often, however, they yield different answers. As an example, consider the results obtained by Yorke and Longden (2005) when they analysed the data from the 2004 *Sunday Times University Guide*. These consisted of the values of seven indicators from a total of 119 institutions. As in the analysis of the data from the 2007 *Sunday Times University Guide* in the present report, the correlation coefficients among the variables were all positive, and in some cases they were very high, suggesting a good deal of overlap in what was being measured.

Figure 4 shows a scree plot obtained from the matrix of correlation coefficients published in Yorke and Longden's report (2005). Once again, the solid line shows the eigenvalues of the principal components obtained from their data. Yorke and Longden extracted three components that explained 86% of the variation in the data. They did not say why they had chosen to extract three components, but the visual scree test might have led them to this conclusion. However, the eigenvalues of the first three components are 4.151, 1.069 and 0.870; according to the eigenvalues-greater-than-one rule, only two components should have been extracted. Moreover, the second eigenvalue is only marginally greater than one and, as mentioned above, even this rule tends to overestimate the true number of components in a dataset.

The dotted line in Figure 4 shows the results of a parallel analysis of 1,000 random correlation matrices using the program written by O'Connor (2000). Once again, the eigenvalues of the first few principal components are actually greater than one. The first principal component is the only one with an obtained eigenvalue that is greater than what would be expected from purely random data, and this implies that Yorke and Longden should actually have extracted only one principal component from the data in the 2004 *Sunday Times University Guide*. In the present report, the same conclusion is reached with regard to the 2007 *Sunday Times University Guide*. These results suggest that this particular league table is somewhat more homogeneous than Yorke and Longden's account might lead one to believe.

Figure 4: Scree plot for Yorke and Longden's (2005) data



Returning to the analysis of handedness, Table 3 shows the results of extracting just one principal component from the data in Table 2. The first column of numbers shows the loadings of each of the variables on the principal component. These can be effectively understood as the correlation coefficients between the observed variables and the hypothetical trait. All of the variables show substantial relationships with the trait in question, supporting the idea that it can be interpreted as a global concept of handedness. The second column of numbers shows the coefficients or loadings that would be used to calculate a person's score on a new variable of handedness from the standard scores of their responses concerning individual activities. (Their scores on this new variable would also be standard scores.) On both counts, all of the individual activities make a contribution, but writing and throwing appear to constitute the most important measures of handedness.

Table 3: Component loadings and component score coefficients for the handedness data

	Loadings	Coefficients
1. Writing	.913	.222
2. Throwing	.895	.218
3. Cutting with scissors	.777	.189
4. Playing with a racket or bat	.697	.170
5. Brushing one's teeth	.807	.196
6. Striking a match	.856	.208

In other situations, the pattern of correlation coefficients among a set of variables will imply the existence of two or more underlying traits, and there are different methods for identifying these. It may then be necessary to transform (or 'rotate') these components to achieve the most meaningful interpretation; depending upon the techniques used, this can result in rotated components that are either orthogonal (i.e. uncorrelated with one another) or oblique (i.e. correlated with one another). Yorke (1997) used orthogonal rotation, but this gives rise to uncorrelated components purely as a statistical artefact. In contrast, oblique rotation gives rise to components that are correlated with one another, but it also subsumes the possibility that the correlation between them is relatively small or zero (in other words, that the components are essentially uncorrelated with one another).

For instance, inventories on approaches to studying often ask students to indicate the extent to which they agree or disagree with various statements as descriptions of their own studying behaviour. The use of factor analysis may produce two underlying factors that represent the extent to which students use a deep approach (based upon understanding the meaning of their course materials) or a surface approach (based upon being able to regurgitate those materials for the purposes of assessment). Perhaps rather counter-intuitively, these two factors often prove to be relatively uncorrelated with one another. This means that the *same* students can score high or low on *both* factors, and trying to dissuade students from using a surface approach will not in itself ensure that they use a deep approach. Of the five league tables we analysed, three yielded just one principal component, but two yielded two principal components. In both cases, the two components proved to be essentially uncorrelated with one another.

References

- Cliff, N. (1988) 'The eigenvalues-greater-than-one rule and the reliability of components', *Psychological Bulletin*, 103(2), pp 276-279.
- O'Connor, B.P. (2000) 'SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test', *Behaviour Research Methods, Instruments, and Computers*, 32(3), pp 396-402.

- Richardson, J.T.E. (1978) 'A factor analysis of self-reported handedness'. *Neuropsychologia*, 16, pp 747-748.
- Yorke, M. (1997) 'A good league table guide?', *Quality Assurance in Education*, 5(2), pp 61-72.
- Yorke, M. (1998) 'The Times' "league table" of universities, 1997: a statistical appraisal', *Quality Assurance in Education*, 6(1), pp 58-60.
- Yorke, M. and Longden, B. (2005) *Significant Figures – Performance Indicators and 'League Tables'*, London: Standing Conference of Principals (now GuildHE)
<http://www.scop.ac.uk/UploadFolder/SCOPsigfigfinalprint2.pdf>