Exploratory factor analysis

17

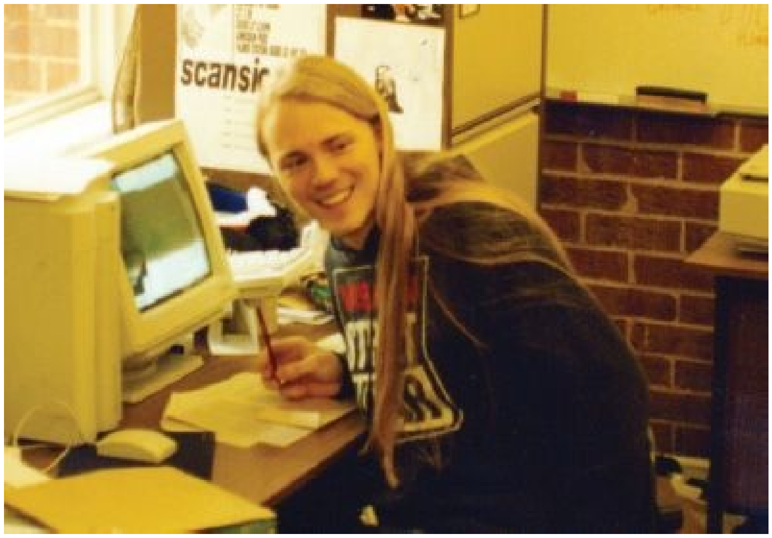


FIGURE 17.1
In my office during my Ph.D., probably preparing some teaching – I had quite long hair back then because it hadn't started falling out at that point

17.1. What will this chapter tell me? ①

Having failed to become a rock star, I went to university and eventually ended up doing a Ph.D. (in Psychology) at the University of Sussex. Like many postgraduates, I taught to survive. I was allocated to second-year undergraduate statistics. I was very shy at the time, and I didn't have a clue about statistics, so standing in front of a room full of strangers and talking to them about ANOVA was about as appealing as dislocating my knees and running a marathon. I obsessively prepared for my first session so that it would go well; I created handouts, I invented examples, I rehearsed what I would say. I went in terrified but knowing that if preparation was any predictor of success then I would be OK. About half way through one of the students rose majestically from her chair. An aura of bright white light surrounded her and she appeared to me as though walking through dry ice. I guessed that she had been chosen by her peers to impart a message of gratitude for the hours of preparation I had done and the skill with which I was unclouding their brains of statistical mysteries. She stopped inches away from me. She looked into my eyes and mine raced around the floor looking for the reassurance of my shoelaces. 'No one in this room has a rabbit[1] clue what you're going on about', she spat before storming out. Scales have not been invented yet to measure how much I wished I'd run the dislocated-knees marathon that morning. To this day I have intrusive thoughts about students in my lectures walking zombie-like towards the front of the lecture theatre chanting 'No one knows what you're going on about' before devouring my brain in a rabid feeding frenzy.

The aftermath of this trauma is that I threw myself into trying to be the best teacher in the universe. I wrote detailed handouts and started using wacky examples. Based on these I was signed up by a publisher to write a book. This book. At the age of 23 I didn't realize that this was academic suicide (really, textbooks take a long time to write and they are not at all valued compared to research articles), and I also didn't realize the emotional pain I was about to inflict on myself. I soon discovered that writing a statistics book was like doing a factor analysis: in factor analysis we take a lot of information (variables) and SPSS effortlessly reduces this mass of confusion into a simple message (fewer variables). SPSS does this in a few seconds. Similarly, my younger self took a mass of information about statistics that I didn't understand and filtered it down into a simple message that I *could* understand: I became a living, breathing factor analysis … except that, unlike SPSS, it took me two years and some considerable effort.

17.2. When to use factor analysis ②

In science we often need to measure things that cannot be measured directly (so-called **latent variables**). For example, management researchers might be interested in measuring 'burnout', which is when someone who has been working very hard on a project (a book, for example) for a prolonged period of time suddenly finds himself devoid of motivation, inspiration, and wants to repeatedly headbutt their computer, screaming 'please, Mike, unlock the door, let me out of the basement, I need to feel the soft warmth of sunlight on my skin'. You can't measure burnout directly: it has many facets. However, you can measure different aspects of burnout: you could get some idea of motivation, stress levels, whether the person has any new ideas and so on. Having done this, it would be helpful to know whether these facets reflect a single variable. Put another way, are these different measures driven by the same underlying variable?

This chapter explores **factor analysis** and **principal component analysis (PCA)** – techniques for identifying clusters of variables. These techniques have three main uses: (1) to understand the structure of a set of variables (e.g., Spearman and Thurstone used factor analysis to try to understand the structure of the latent variable 'intelligence'); (2) to construct a questionnaire to measure an underlying variable (e.g., you might design a questionnaire to measure burnout); and (3) to reduce a data set to a more manageable size while retaining as much of the original information as possible (e.g., factor analysis can be used to solve the problem of multicollinearity that we discovered in Chapter 8 by combining variables that are collinear).

There are numerous examples of the use of factor analysis in science. Most readers will be familiar with the extroversion−introversion and neuroticism traits measured by Eysenck (1953). Most other personality questionnaires are also based on factor analysis – notably Cattell's (1966a) 16 personality factors questionnaire – and these inventories are frequently used for recruiting purposes in industry (and even by some religious groups). Economists, for example, might also use factor analysis to see whether productivity, profits and workforce can be reduced down to an underlying dimension of company growth, and Jeremy Miles told me of a biochemist who used it to analyse urine samples.

Both factor analysis and PCA aim to reduce a set of variables into a smaller set of dimensions (called 'factors' in factor analysis and 'components' in PCA). To non-statisticians, like me, the differences between a component and a factor are difficult to conceptualize (they are both linear models), and the differences are hidden away in the maths behind the techniques.[2] However, there are important differences between the techniques, which I'll discuss in due course. Most of the practical issues are the same regardless of whether you do factor analysis or PCA, so once the theory is over you can apply any advice I give to either factor analysis or PCA.

17.3. Factors and components ②

If we measure several variables, or ask someone several questions about themselves, the correlation between each pair of variables (or questions) can be arranged in a table (just like the output from a correlation analysis as seen in Chapter 7). This table is sometimes called an *R*-matrix, just to scare you. The diagonal elements of an *R*-matrix are all ones because each variable will correlate perfectly with itself. The off-diagonal elements are the correlation coefficients between pairs of variables, or questions.[3] Factor analysis attempts to achieve parsimony by explaining the maximum amount of *common variance* in a correlation matrix using the smallest number of explanatory constructs. These 'explanatory constructs' are known as **factors** (or *latent variables*) in factor analysis, and they represent clusters variables that correlate highly with each other. PCA tries to explain the maximum amount of *total variance* (not just common variance) in a correlation matrix by transforming the original variables into linear **components**.

Imagine that we wanted to measure different aspects of what might make a person popular. We could administer several measures that we believe tap different aspects of popularity. So, we might measure a person's social skills (**Social Skills**), their selfishness (**Selfish**), how interesting others find them (**Interest**), the proportion of time they spend talking about the other person during a conversation (**Talk1**), the proportion of time they spend talking about themselves (**Talk2**), and their propensity to lie to people (**Liar**). We calculate the correlation coefficients for each pair of variables and create an *R*-matrix. Figure 17.2 shows this matrix. There appear to be two clusters of interrelating variables. First, the amount that someone talks about the other person during a conversation correlates highly with both the level of social skills and how interesting the other finds that person, and social skills correlate well with how interesting others perceive a person to be. These relationships indicate that the better your social skills, the more interesting and talkative you are likely to be. Second, the amount that people talk about themselves within a conversation correlates with how selfish they are and how much they lie. Being selfish also correlates with the degree to which a person tells lies. In short, selfish people are likely to lie and talk about themselves.

	Talk 1	Social Skills	Interest	Talk 2	Selfish	Liar
Talk 1	1.000					
Social Skills	.772	1.000				
Interest	.646	.879	1.000			
Talk 2	.074	−.120	.054	1.000		
Selfish	−.131	.031	−.101	.441	1.000	
Liar	.068	.012	.110	.361	.277	1.000

Factor 1

Factor 2

FIGURE 17.2
An *R*-matrix

Factor analysis and PCA both aim to reduce this *R*-matrix down into a smaller set of dimensions. In factor analysis these dimensions, or factors, are estimated from the data and are believed to reflect constructs that can't be measured directly. In this example, there appear to be two clusters that fit the bill. The first 'factor' seems to relate to general sociability, whereas the second 'factor' seems to relate to the way in which a person treats others socially (we might call it *Consideration*). It might, therefore, be assumed that popularity depends not only on your ability to socialize, but also on whether you are inconsiderate towards others. PCA, in contrast, transforms the data into a set of linear components; it does not estimate unmeasured variables, it just transforms measured ones. Strictly speaking, then, we shouldn't interpret components as unmeasured variables. Despite these differences, both techniques look for variables that correlate highly with a group of other variables, but do not correlate with variables outside of that group.

17.3.1. Graphical representation ②

Factors and components can also be visualized: you can imagine factors as being the axis of a graph along which we plot variables. The coordinates of variables along each axis represent the strength of relationship between that variable and each factor. In an ideal world a variable should have a large coordinate for one of the axes, and small coordinates for any other factors. This scenario would indicate that this particular variable related to only one factor. Variables that have large coordinates on the same axis are assumed to measure different aspects of some common underlying dimension. The coordinate of a variable along a classification axis is known as a **factor loading** (or component loading). The factor loading can be thought of as the Pearson correlation between a factor and a variable (see Jane Superbrain Box 17.1). From what we know about interpreting correlation coefficients (see Section 7.4.2.2) it should be clear that if we square the factor loading we obtain a measure of the substantive importance of a particular variable to a factor.

Figure 17.3 shows such a plot for the popularity data (in which there were only two factors). The first thing to notice is that for both factors, the axis line ranges from −1 to 1, which are the outer limits of a correlation coefficient. The triangles represent the three variables that have high factor loadings (i.e., a strong relationship) with factor 1 (**Sociability**: horizontal axis) but have a low correlation with factor 2 (**Consideration**: vertical axis). Conversely, the circles represent variables that have high factor loadings with consideration but low loadings with sociability. This plot shows what we found in the *R*-matrix: selfishness, the amount a person talks about themselves and their propensity to lie contribute to a factor which could be called consideration of others; and how much a person takes an interest in other people, how interesting they are and their level of social skills contribute to a second factor, sociability. Of course, if a third factor existed within these data it could be represented by a third axis (creating a 3-D graph). If more than three factors exist in a data set, then they cannot all be represented by a 2-D plot.

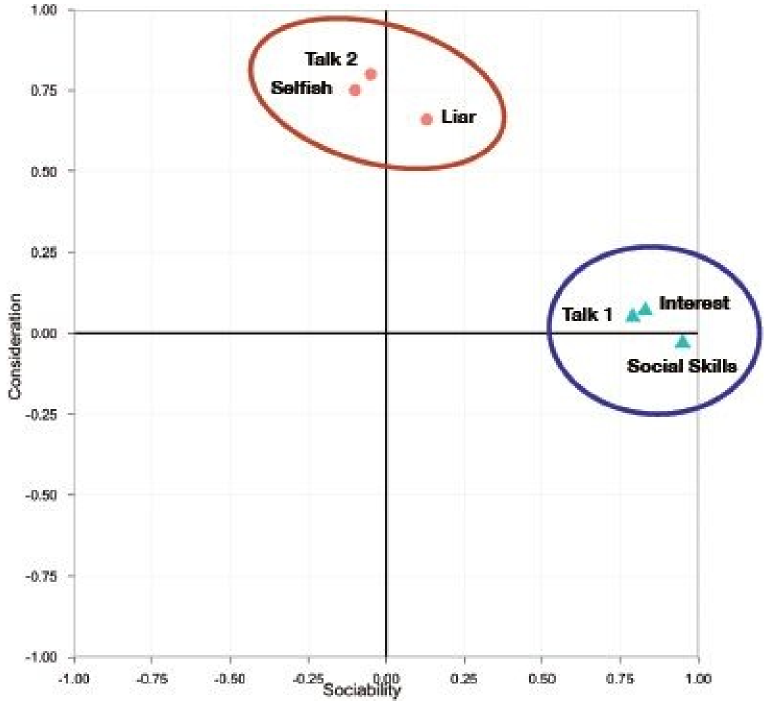


FIGURE 17.3
Example of a factor plot

17.3.2. Mathematical representation ②

The axes in Figure 17.3, which represent factors, are straight lines and any straight line can be described mathematically by a familiar equation.


SELF-TEST What is the equation of a straight line/linear model?

Equation (17.1) reminds us of the equation describing a linear model. A component in PCA can be described in the same way. You'll notice that there is no intercept in the equation because the lines intersect at zero (hence the intercept is zero), and there is also no error term because we are simply transforming the variables. The *b*s in the equation represent the loadings.

SMART ALEX ONLY

$$\begin{aligned} Y_i &= b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni} \\ \text{Component}_i &= b_1\text{Variable}_{1i} + b_2\text{Variable}_{2i} + \cdots + b_n\text{Variable}_{ni} \end{aligned} \quad (17.1)$$

Sticking with our example of popularity, we found that there were two components: general sociability and consideration. We can, therefore, construct an equation that describes each factor in terms of the variables that have been measured. The equations are as follows:

$$\begin{aligned} Y_i &= b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni} \\ \text{Sociability}_i &= b_1\text{Talk1}_i + b_2\text{Social Skills}_i + b_3\text{Interest}_i \\ &\quad + b_4\text{Talk2}_i + b_5\text{Selfish}_i + b_6\text{Liar}_i \\ \text{Consideration}_i &= b_1\text{Talk1}_i + b_2\text{Social Skills}_i + b_3\text{Interest}_i \\ &\quad + b_4\text{Talk2}_i + b_5\text{Selfish}_i + b_6\text{Liar}_i \end{aligned} \quad (17.2)$$

First, notice that the equations are identical in form: they both include all of the variables that were measured. However, the values of *b* in the two equations will be different (depending on the relative importance of each variable to the particular component). In fact, we can replace each value of *b* with the coordinate of that variable on the graph in Figure 17.3 (i.e., replace the values of *b* with the factor loadings). The resulting equations are as follows:

$$\begin{aligned} Y_i &= b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni} \\ \text{Sociability}_i &= 0.87\text{Talk1}_i + 0.96\text{Social Skills}_i + 0.92\text{Interest}_i + 0.00\text{Talk2}_i \\ &\quad - 0.10\text{Selfish}_i + 0.09\text{Liar}_i \\ \text{Consideration}_i &= 0.01\text{Talk1}_i - 0.03\text{Social Skills}_i + 0.04\text{Interest}_i + 0.82\text{Talk2}_i \\ &\quad + 0.75\text{Selfish}_i + 0.70\text{Liar}_i \end{aligned} \quad (17.3)$$

Notice that, for the **Sociability** component, the values of *b* are high for **Talk1**, **Social Skills** and **Interest**. For the remaining variables (**Talk2**, **Selfish** and **Liar**) the values of *b* are very low (close to 0). This tells us that three of the variables are very important for that component (the ones with high values of *b*) and three are very unimportant (the ones with low values of *b*). We saw that this point is true because of the way that three variables clustered highly on the factor plot (Figure 17.3). The point to take on board here is that the factor plot and these equations represent the same thing: the factor loadings in the plot are simply the *b*-values in these equations. For the second factor, **Consideration**, the opposite pattern can be seen: **Talk2**, **Selfish** and **Liar** all have high values of *b*, whereas the remaining three variables have *b*-values close to 0. In an ideal world, variables would have very high *b*-values for one component and very low *b*-values for all other components.

The factors in factor analysis are not represented in quite the same way as components. Equation (17.4) shows how a factor is defined: the Greek letters represent matrices containing numbers. If we put the Greek letters through Andy's magical translation machine then we can stop worrying about what the matrices contain and focus on what they represent. In factor analysis, scores on the measured variables are predicted from the means of those variables plus a person's scores on the **common factors** (i.e., factors that explain the correlations between variables) multiplied by their factor loadings, plus scores on any **unique factors** within the data (factors that cannot explain the correlations between variables).

$$\begin{aligned} x &= \mu + \Lambda\xi + \delta \\ \text{Variables} &= \text{Variable Means} + (\text{Loadings} \times \text{Common Factor}) + \text{Unique Factor} \end{aligned} \quad (17.4)$$

In a sense, the factor analysis model flips PCA on its head: in PCA we predict components from the measured variables, but in factor analysis we predict the measured variables from the underlying factors. For example, psychologists are usually interested in factors, because they're interested in how the stuff going on inside people's heads (the latent variables) affects how they answer the questions (the measured variables). The other big difference is that, unlike PCA, factor analysis contains an error term (δ is made up of both scores on unique factors and measurement error). The fact that PCA assumes that there is no measurement error upsets a lot of people who use factor analysis.

Both factor analysis and PCA are linear models in which loadings are used as weights. In both cases, these loadings can be expressed as a matrix in which the columns represent each factor and the rows represent the loadings of each variable on each factor. For the popularity data this matrix would have two columns (one for each factor) and six rows (one for each variable). This matrix, Λ, can be seen below. It is called the **factor matrix** or **component matrix** (if doing principal component analysis) – see Jane Superbrain Box 17.1 to find out about the different forms of this matrix. Try relating the elements to the loadings in equation (17.3) to give you an idea of what this matrix represents (in the case of PCA). For example, the top row represents the first variable, **Talk1**, which had a loading of .87 for the first factor (**Sociability**) and a loading of .01 for the second factor (**Consideration**).

$$\Lambda = \begin{pmatrix} 0.87 & 0.01 \\ 0.96 & -0.03 \\ 0.92 & 0.04 \\ 0.00 & 0.82 \\ -0.10 & 0.75 \\ 0.09 & 0.70 \end{pmatrix}$$

The major assumption in factor analysis (but not PCA) is that these algebraic factors represent real-world dimensions, the nature of which must be *guessed at* by inspecting which variables have high loads on the same factor. So, psychologists might believe that factors represent dimensions of the psyche, education researchers might believe they represent abilities, and sociologists might believe they represent races or social classes. However, it is an extremely contentious point: some believe that the dimensions derived from factor analysis are real only in the statistical sense – and are real-world fictions.

EVERYBODY

17.3.3. Factor scores ②

A factor can be described in terms of the variables measured and their relative importance for that factor. Therefore, having discovered which factors exist, and estimated the equation that describes them, it should be possible to estimate a person's score on a factor, based on their scores for the constituent variables; these are known as **factor scores** (or *component scores* in PCA). For example, if we wanted to derive a sociability score for a particular person after PCA, we could place their scores on the various measures into equation (17.3). This method is known as a *weighted average* and is rarely used because it is overly simplistic, but it is the easiest way to explain the principle. For example, imagine our six personality measures range from 1 to 10 and that someone scored the following: **Talk1** (4), **Social Skills** (9), **Interest** (8), **Talk2** (6), **Selfish** (8), and **Liar** (6). We could plug these values into equation (17.3) to get a score for this person's sociability and their consideration to others (see equation (17.5)). The resulting scores of 19.22 and 15.21 reflect the degree to which this person is sociable and their inconsideration towards others, respectively. This person scores higher on sociability than inconsideration. However, the scales of measurement used will influence the resulting scores, and if different variables use different measurement scales, then factor scores for different factors cannot be compared. As such, this method of calculating factor scores is poor and more sophisticated methods are usually used:

$$
\begin{aligned}
\text{Sociability}_i &= 0.87\text{Talk1}_i + 0.96\text{Social Skills}_i + 0.92\text{Interest}_i + 0.00\text{Talk2}_i \\
&\quad - 0.10\text{Selfish}_i + 0.09\text{Liar}_i \\
\text{Sociability}_i &= (0.87 \times 4) + (0.96 \times 9) + (0.92 \times 8) + (0.00 \times 6) \\
&\quad - (0.10 \times 8) + (0.09 \times 6) \\
&= 19.22 \\
\text{Consideration}_i &= 0.01\text{Talk1}_i - 0.03\text{Social Skills}_i + 0.04\text{Interest}_i + 0.82\text{Talk2}_i \\
&\quad + 0.75\text{Selfish}_i + 0.70\text{Liar}_i \\
\text{Consideration}_i &= (0.01 \times 4) - (0.03 \times 9) + (0.04 \times 8) + (0.82 \times 6) \\
&\quad + (0.75 \times 8) + (0.70 \times 6) \\
&= 15.21
\end{aligned}
\tag{17.5}
$$

JANE SUPERBRAIN 17.1

What's the difference between a pattern matrix and a structure matrix? ③

So far I've been a bit vague about factor loadings. Sometimes I've said that these loadings can be thought of as the correlation between a variable and a given factor, then at other times I've described these loadings in terms of regression coefficients (*b*). Broadly speaking, both correlation coefficients and regression coefficients represent the relationship between a variable and linear model, so my vagueness might not be the evidence of buffoonery that it initially seems. The take-home message is that factor loadings tell us about the relative contribution that a variable makes to a factor. As long as you understand that much, you'll be OK.

However, the factor loadings in a given analysis can be both correlation coefficients and regression coefficients. In a few sections' time we'll discover that the interpretation of factor analysis is helped greatly by a technique known as *rotation*. Without going into details, there are two types: orthogonal and oblique rotation (see Section 17.4.6). When orthogonal rotation is used, any underlying factors are assumed to be independent, and the factor loading *is* the correlation between the factor and the variable, but it is also the regression coefficient. Put another way, the values of the correlation coefficients are the same as the values of the regression coefficients. However, there are situations in which the underlying factors are assumed to be related or correlated to each other. In these situations, oblique rotation is used and the resulting correlations between variables and factors will differ from the corresponding regression coefficients. In this case, there are, in effect, two different sets of factor loadings: the correlation coefficients between each variable and factor (which are put in the factor structure matrix) and the regression coefficients for each variable on each factor (which are put in the factor pattern matrix). These coefficients can have quite different interpretations (see Graham, Guthrie, & Thompson, 2003).

17.3.3.1. The regression method ④

There are several sophisticated techniques for calculating factor scores that use factor score coefficients as weights rather than using the factor loadings. Factor score coefficients can be calculated in several ways. The simplest way is the regression method. In this method the factor loadings are adjusted to take account of the initial correlations between variables; in doing so, differences in units of measurement and variable variances are stabilized.

To obtain the matrix of factor score coefficients (B) we multiply the matrix of factor loadings by the inverse (R^{-1}) of the original correlation or R-matrix (this is the same process that is used to estimate the *b*s in ordinary regression). You might remember from the previous chapter that matrices cannot be divided (see Section 16.4.4.1). Therefore, the equivalent of dividing by a matrix is to multiply by the inverse of that matrix. Conceptually speaking, then, by multiplying the matrix of factor loadings by the inverse of the correlation matrix we are dividing the factor loadings by the correlation coefficients. The resulting factor score matrix represents the relationship between each variable and each factor, taking into account the original relationships between pairs of variables. As such, this matrix represents a purer measure of the *unique* relationship between variables and factors.

The regression technique ensures that the resulting factor scores have a mean of 0 and a variance equal to the squared multiple correlation between the estimated factor scores and the true factor values. However, the downside is that the scores can correlate not only with factors other than the one on which they are based, but also with other factor *scores* from a different orthogonal factor.

OLIVER TWISTED

Please Sir, can I have some more … matrix algebra?

‘*The Matrix* …’, enthuses Oliver, ‘… that was a good film. I want to dress in black and glide through the air as though time has stood still. Maybe the matrix of factor scores is as cool as the film.’ I think you might be disappointed, Oliver, but we’ll give it a shot. The matrix calculations of factor score coefficients for this example are detailed in the additional material for this chapter on the companion website. Be afraid, be very afraid …

17.3.3.2. Other methods ②

To overcome the problems associated with the regression technique, two adjustments have been proposed: the Bartlett method and the **Anderson–Rubin method**. The Bartlett method produces scores that are unbiased and that correlate only with their own factor. The mean and standard deviation of the scores is the same as for the regression method. However, factor scores can still correlate with each other. The Anderson-Rubin method is a modification of the Bartlett method that produces factor scores that are uncorrelated and standardized (they have a mean of 0 and a standard deviation of 1). Tabachnick and Fidell (2012) conclude that the Anderson–Rubin method is best when uncorrelated scores are required but that the regression method is preferred in other circumstances simply because it is most easily understood. Although it isn't important that you understand the maths behind any of the methods, it is important that you understand what the factor scores represent: namely, a composite score for each individual on a particular factor.

17.3.3.3. Uses of factor scores ②

There are several uses of factor scores. First, if the purpose of the factor analysis is to reduce a large set of data to a smaller subset of measurement variables, then the factor scores tell us an individual's score on this subset of measures. Therefore, any further analysis can be carried out on the factor scores rather than the original data. For example, we could carry out a *t*-test to see whether females are significantly more sociable than males using the factor scores for sociability. A second use is in overcoming collinearity problems in regression. If, following a multiple regression analysis, we have identified sources of multicollinearity then the interpretation of the analysis is compromised (see Section 8.5.3). In this situation, we can carry out a PCA on the predictor variables to reduce them to a subset of uncorrelated factors. The variables causing the multicollinearity will combine to form a component. If we then rerun the regression but using the component scores as predictor variables then the problem of multicollinearity should vanish (because the variables are now combined into a single component). There are ways in which we can ensure that the components are uncorrelated (one way is to use the Anderson–Rubin method – see above). By using uncorrelated component scores as predictors in the regression we can be confident that there will be no correlation between predictors – hence, no multicollinearity.

17.4. Discovering factors ②

By now, you should have some grasp of what a factor is and what a component is, so we will now delve into how to find or estimate these mythical beasts.

17.4.1. Choosing a method ②

There are several methods for unearthing factors in your data. The method you choose will depend on what you hope to do with the analysis. Tinsley and Tinsley (1987) give an excellent account of the different methods available. There are two things to consider: whether you want to generalize the findings from your sample to a population and whether you are exploring your data or testing a specific hypothesis. This chapter describes techniques for exploring data using factor analysis. Testing hypotheses about the structures of latent variables and their relationships to each other requires considerable complexity and can be done with computer programs such as SPSS's sister package, AMOS. Those interested in hypothesis testing techniques (known as **confirmatory factor analysis**) are advised to read Pedhazur and Schmelkin (1991: Chapter 23) for an introduction.

Assuming we want to explore our data, we then need to consider whether we want to apply our findings to the sample collected (descriptive method) or to generalize our findings to a population (inferential methods). When factor analysis was originally developed it was assumed that it would be used to explore data to generate future hypotheses. As such, it was assumed that the technique would be applied to the entire population of interest. Therefore, certain techniques assume that the sample used is the population, and so results cannot be extrapolated beyond that particular sample. Principal component analysis is an example of these techniques, as are principal factors analysis (*principal axis factoring*) and image covariance analysis (*image factoring*). Of these, principal component analysis and principal factors analysis are the preferred methods and usually result in similar solutions (see Section 17.4.3). When these methods are used, conclusions are restricted to the sample collected and generalization of the results can be achieved only if analysis using different samples reveals the same factor structure (i.e., cross-validation).

Another approach is to assume that participants are randomly selected and that the variables measured constitute the population of variables in which we're interested. By assuming this, it is possible to generalize from the sample participants to a larger population, but with the caveat that any findings hold true only for the set of variables measured (because we've assumed this set constitutes the entire population of variables). Techniques in this category include the *maximum-likelihood method* (see Harman, 1976) and Kaiser's *alpha factoring*. The choice of method depends largely on what generalizations, if any, you want to make from your data.

17.4.2. Communality ②

The idea of what variance is and how it is calculated should, by now, be an old friend with whom you enjoy tea and biscuits (if not, see Chapter 2). The total variance for a particular variable in the *R*-matrix will have two components: some of it will be shared with other variables or measures (**common variance**) and some of it will be specific to that measure (**unique variance**). We tend to use the term *unique variance* to refer to variance that can be reliably attributed to only one measure. However, there is also variance that is specific to one measure but not reliably so; this variance is called error or **random variance**. The proportion of common variance present in a variable is known as the **communality**. As such, a variable that has no unique variance (or random variance) would have a communality of 1; a variable that shares none of its variance with any other variable would have a communality of 0.

In factor analysis we are interested in finding common underlying dimensions within the data and so we are primarily interested only in the common variance. Therefore, we need to know how much of the variance present in our data is common variance. This presents us with a logical impasse: to do the factor analysis we need to know the proportion of common variance present in the data, yet the only way to find out the extent of the common variance is by carrying out a factor analysis! There are two ways to approach this problem. The first is to assume that all of the variance is common variance: we assume that the communality of every variable is 1. By making this assumption we merely transpose our original data into constituent linear components. This procedure is PCA. Remember that I said earlier that PCA assumes no measurement error? Well, by setting the communalities to 1, we are assuming that all variance is common variance (there is no random variance at all).

The second approach is to estimate the amount of common variance by estimating communality values for each variable. There are various methods of estimating communalities but the most widely used (including **alpha factoring**) is to use the squared multiple correlation (SMC) of each variable with all others. So, for the popularity data, imagine you ran a multiple regression using one measure (**Selfish**) as the outcome and the other five measures as predictors: the resulting multiple R^2 (see Section 8.2.4) would be used as an estimate of the communality for the variable **Selfish**. This second approach is used in factor analysis. These estimates allow the factor analysis to be done. Once the underlying factors have been extracted, new communalities can be calculated that represent the multiple correlation between each variable and the factors extracted. Therefore, the communality is a measure of the proportion of variance explained by the extracted factors.

17.4.3. Factor analysis or PCA? ②

I have just explained that there are two approaches to locating underlying dimensions of a data set: factor analysis and principal component analysis. These techniques differ in the communality estimates that are used. As I have hinted before, factor analysis derives a mathematical model from which factors are estimated, whereas PCA decomposes the original data into a set of linear variates (see Dunteman, 1989, Chapter 8, for more detail on the differences between the procedures). As such, only factor analysis can estimate the underlying factors, and it relies on various assumptions for these estimates to be accurate. PCA is concerned only with establishing which linear components exist within the data and how a particular variable might contribute to that component.

Based on an extensive literature review, Guadagnoli and Velicer (1988) concluded that the solutions generated from PCA differ little from those derived from factor-analytic techniques. In reality, with 30 or more variables and communalities greater than 0.7 for all variables, different solutions are unlikely; however, with fewer than 20 variables and any low communalities (< 0.4) differences can occur (Stevens, 2002).

The flip side of this argument is eloquently described by Cliff (1987) who observed that proponents of factor analysis 'insist that components analysis is at best a common factor analysis with some error added and at worst an unrecognizable hodgepodge of things from which nothing can be determined' (p. 349). Indeed, feeling is strong on this issue, with some arguing that when PCA is used it should not be described as a factor analysis (oops!) and that you should not impute substantive meaning to the resulting components. Ultimately, as I hope to have made clear, they are doing slightly different things.

17.4.4. Theory behind PCA ③

The theory behind factor analysis is, frankly, a bit of an arse; an arse tattooed with matrix algebra. No-one wants to look at matrix algebra when they're admiring an arse, so we'll look at the squeezable buttocks of PCA instead. Principal component analysis works in a very similar way to MANOVA and discriminant function analysis (see Chapter 16). In MANOVA, various sum of squares and crossproduct matrices were calculated that contained information about the relationships between dependent variables. I mentioned before that these SSCP matrices can be converted to variance–covariance matrices, which represent the same information but in averaged form (i.e., taking account of the number of observations). I also pointed out that by dividing each element by the relevant standard deviation the variance–covariance matrices becomes standardized. The result is a correlation matrix. In PCA we usually deal with correlation matrices (although it is possible to analyse a variance–covariance matrix too), and my point is that this matrix represents the same information as an SSCP matrix in MANOVA.

SMART ALEX ONLY

In MANOVA, because we were comparing groups we ended up looking at the variates or components of the SSCP matrix that represented the ratio of the model variance to the error variance. These variates were linear dimensions that separated the groups tested, and we saw that the dependent variables mapped onto these underlying components. In short, we looked at whether the groups could be separated by some linear combination of the dependent variables. These variates were found by calculating the eigenvectors of the SSCP. The number of variates obtained was the smaller of p (the number of dependent variables) or $k - 1$ (where k is the number of groups).

In PCA we do much the same thing but using the overall correlation matrix (because we're not interested in comparing groups of scores). To simplify things a little, we take a correlation matrix and calculate the variates. There are no groups of observations, and so the number of variates calculated will always equal the number of variables measured (p). The variates are described, as for MANOVA, by the eigenvectors associated with the correlation matrix. The elements of the eigenvectors are the weights of each variable on the variate. These values are the loadings described earlier (i.e., the b-values in equation (16.5)). The largest eigenvalue associated with each of the eigenvectors provides a single indicator of the substantive importance of each component. The basic idea is that we retain components with relatively large eigenvalues and ignore those with relatively small eigenvalues.

Factor analysis works differently, but there are similarities. Rather than using the correlation matrix, factor analysis starts by estimating the communalities between variables using the SMC (as described earlier). It then replaces the diagonal of the correlation matrix (the 1s) with these estimates. Then the eigenvectors and associated eigenvalues of this matrix are computed. Again, these eigenvalues tell us about the substantive importance of the factors, and based on them a decision is made about how many factors to retain. Loadings and communalities are then estimated using only the retained factors.

EVERYBODY

17.4.5. Factor extraction: eigenvalues and the scree plot ②

In both PCA and factor analysis, not all factors are retained. The process of deciding how many factors to keep is called *extraction*. I mentioned above that eigenvalues associated with a variate indicate the substantive importance of that factor. Therefore, it is logical to retain only factors with large eigenvalues. This section looks at how we determine whether an eigenvalue is large enough to represent a meaningful factor.

Cattell (1966b) suggested plotting each eigenvalue (*Y*-axis) against the factor with which it is associated (*X*-axis). This graph is known as a **scree plot** (because it looks like a rock face with a pile of debris, or scree, at the bottom). I mentioned earlier that it is possible to obtain as many factors as there are variables and that each has an associated eigenvalue. By graphing the eigenvalues, the relative importance of each factor becomes apparent. Typically there will be a few factors with quite high eigenvalues, and many factors with relatively low eigenvalues, and so this graph has a very characteristic shape: there is a sharp descent in the curve followed by a tailing off (see Figure 17.4). The point of inflexion is where the slope of the line changes dramatically, and Cattell (1966b) suggested using this point as the cut-off for retaining factors. In Figure 17.4, imagine drawing two straight lines (the red dashed lines), one summarizing the vertical part of the plot and the other summarizing the horizontal part. The point of inflexion is the data point at which these two lines meet. You retain only factors to the left of the point of inflexion (and do not include the factor at the point of inflexion itself),[4] so in both examples in Figure 17.4 we would extract two factors because the point of inflexion occurs at the third data point (factor). With a sample of more than 200 participants, the scree plot provides a fairly reliable criterion for factor selection (Stevens, 2002).

Although scree plots are very useful, Kaiser (1960) recommended retaining all factors with eigenvalues greater than 1. This criterion is based on the idea that the eigenvalues represent the amount of variation explained by a factor and that an eigenvalue of 1 represents a substantial amount of variation. Jolliffe (1972, 1986) reports that **Kaiser's criterion** is too strict and suggested retaining all factors with eigenvalues more than 0.7. The difference between how many factors are retained using Kaiser's methods compared to Jolliffe's can be dramatic.

You might well wonder how the methods compare. Generally speaking, Kaiser's criterion overestimates the number of factors to retain (see Jane Superbrain Box 17.2), but there is some evidence that it is accurate when the number of variables is less than 30 and the resulting communalities (after extraction) are all greater than 0.7. Kaiser's criterion can also be accurate when the sample size exceeds 250 and the average communality is greater than or equal to 0.6. In any other circumstances you are best advised to use a scree plot, provided the sample size is greater than 200 (see Stevens, 2002, for more detail). By default, SPSS uses Kaiser's criterion to extract factors. Therefore, if you use the scree plot to determine how many factors are retained you may have to rerun the analysis specifying that SPSS extracts the number of factors you require.

As is often the case in statistics, the three criteria often provide different answers. In these situations the communalities of the factors need to be considered. Remember that communalities represent the common variance: if the values are 1 then all common variance is accounted for, and if the values are 0 then no common variance is accounted for. In both PCA and factor analysis we determine how many factors/components to extract and then re-estimate the communalities. The factors we retain will not explain all of the variance in the data (because we have discarded some information) and so the communalities after extraction will always be less than 1. The factors retained do not map perfectly onto the original variables – they merely reflect the common variance present in the data. If the communalities represent a loss of information then they are important statistics. The closer the communalities are to 1, the better our factors are at explaining the original data. It is logical that the more factors retained, the greater the communalities will be (because less information is discarded); therefore, the communalities are good indices of whether too few factors have been retained. In fact, with generalized least-squares factor analysis and maximum-likelihood factor analysis you can get a statistical measure of the goodness of fit of the factor solution (see the next chapter for more on goodness-of-fit tests). This basically measures the proportion of variance that the factor solution explains (so can be thought of as comparing communalities before and after extraction).

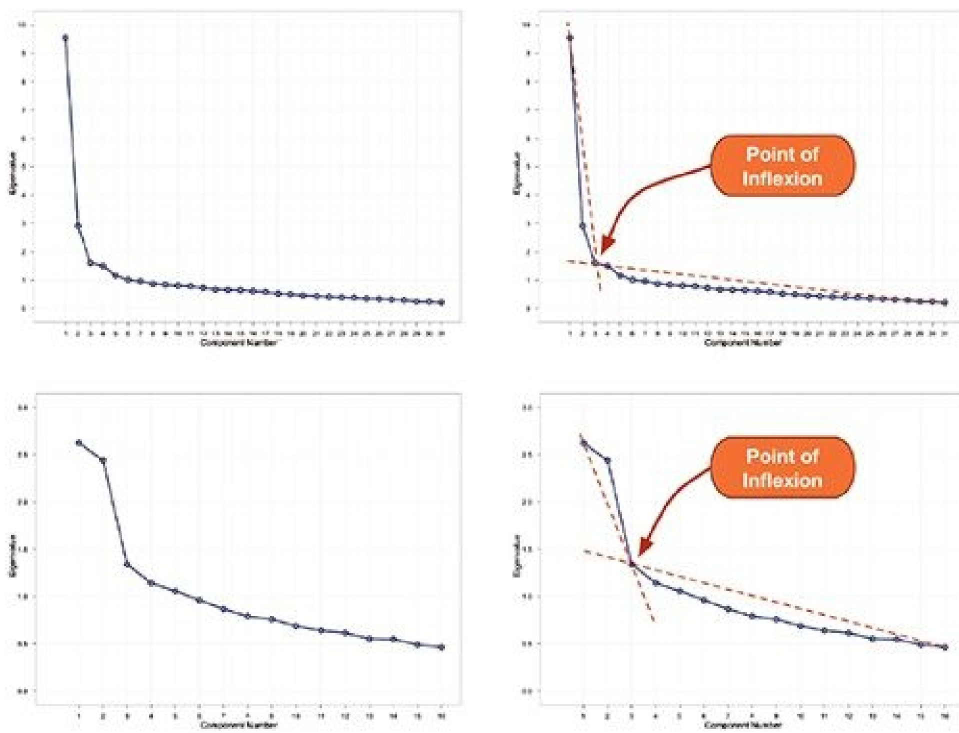


FIGURE 17.4 Examples of scree plots for data that probably have two underlying factors

As a final word of advice, your decision on how many factors to extract will depend also on why you're doing the analysis; for example, if you're trying to overcome multicollinearity problems in regression, then it might be better to extract too many factors than too few.

17.4.6. Improving interpretation: factor rotation ③

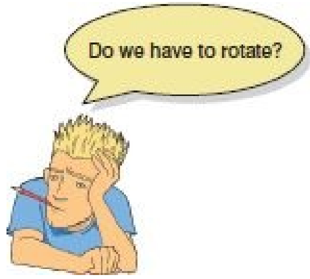
Once factors have been extracted, it is possible to calculate the degree to which variables load on these factors (i.e., calculate the loadings for each variable on each factor). Generally, you will find that most variables have high loadings on the most important factor and small loadings on all other factors. This characteristic makes interpretation difficult, and so a technique called factor **rotation** is used to discriminate between factors. If we visualize our factors as an axis along which variables can be plotted, then factor rotation effectively rotates these axes such that variables are loaded maximally to only one factor. Figure 17.5 demonstrates how this process works using an example in which there are only two factors. Imagine that a sociologist was interested in classifying university lecturers as a demographic group. She discovered that two underlying dimensions best describe this group: alcoholism and achievement (go to any academic conference and you'll see why I chose these dimensions). The first factor, alcoholism, has a cluster of variables associated with it (green circles), and these could be measures such as the number of units drunk in a week, dependency and obsessive personality. The second factor, achievement, also has a cluster of variables associated with it (red circles) and these could be measures relating to salary, job status and number of research publications. Initially, the full lines represent the factors, and by looking at the coordinates it should be clear that the red circles have high loadings for factor 2 (they are a long way up this axis) and medium loadings for factor 1 (they are not very far up this axis). Conversely, the green circles have high loadings for factor 1 and medium loadings for factor 2. By rotating the axes (dashed lines), we ensure that both clusters of variables are intersected by the factor to which they relate most. So, after rotation, the loadings of the variables are maximized on one factor (the factor that intersects the cluster) and minimized on the remaining factor(s). If an axis passes through a cluster of variables, then these variables will have a loading of approximately zero on the opposite axis. If this idea is confusing, then look at Figure 17.5 and think about the values of the coordinates before and after rotation (this is best achieved by turning the book when you look at the rotated axes).

JANE SUPERBRAIN 17.2

How many factors do I retain? ③

There are fundamental problems with Kaiser's criterion (Nunnally & Bernstein, 1994). For one thing, an eigenvalue of 1 means different things in different analyses: with 100 variables it means that a factor explains 1% of the variance, but with 10 variables it means that a factor explains 10% of the variance. Clearly, these two situations are very different and a single rule that covers both is inappropriate. An eigenvalue of 1 also means only that the factor explains as much variance as a variable, which rather defeats the original intention of the analysis to reduce variables down to 'more substantive' underlying factors. Consequently, Kaiser's criterion often overestimates the number of factors. By this argument Jolliffe's criterion is even worse (a factor explains less variance than a variable).

There are more complex ways to determine how many factors to retain, but they are not easy to do in SPSS. The best is probably parallel analysis (Horn, 1965). Essentially each eigenvalue (which represents the size of the factor) is compared against an eigenvalue for the corresponding factor in many randomly generated data sets that have the same characteristics as the data being analysed. In doing so, each eigenvalue is compared to an eigenvalue from a data set that has no underlying factors. This is a bit like asking whether our observed factor is bigger than a non-existing factor. Factors that are bigger than their 'random' counterparts are retained. Of parallel analysis, the scree plot and Kaiser's criterion, Kaiser's criterion is, in general, worst and parallel analysis best (Zwick & Velicer, 1986). If you want to do parallel analysis then SPSS syntax is available (O'Connor, 2000) from https://people.ok.ubc.ca/brioconn/nfactors/nfactors.html.



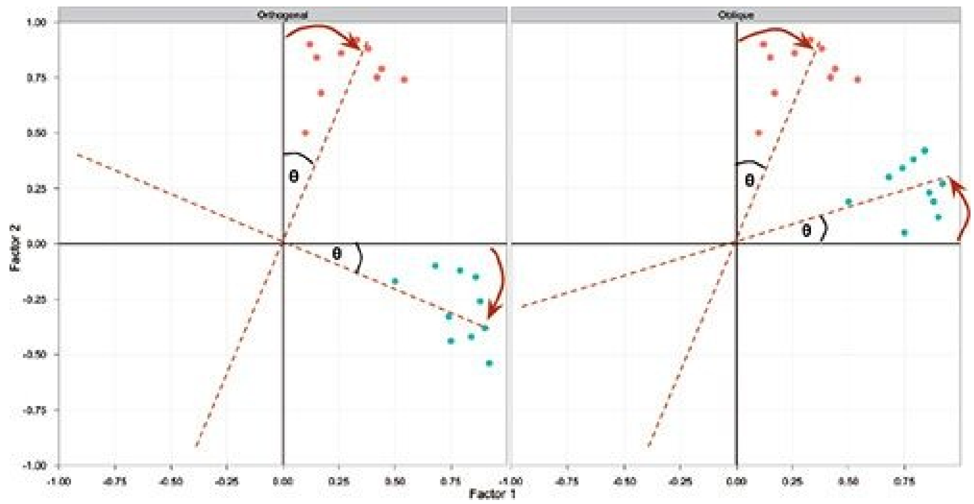


FIGURE 17.5
Schematic representations of factor rotation. The left graph displays orthogonal rotation, whereas the right graph displays oblique rotation (see text for more details). θ is the angle through which the axes are rotated

There are two types of rotation that can be done. The first is **orthogonal rotation**, and the left-hand side of Figure 17.5 represents this method. In Chapter 11 we saw that the term *orthogonal* means 'unrelated', and in this context it means that we rotate factors while keeping them independent, or unrelated. Before rotation, all factors are independent (i.e., they do not correlate at all) and orthogonal rotation ensures that the factors remain uncor-related. That is why in Figure 17.5 the axes are turned while remaining perpendicular.[5] The other form of rotation is **oblique rotation**. The difference with oblique rotation is that the factors are allowed to correlate (hence, the axes of the right-hand diagram of Figure 17.5 do not remain perpendicular).

The choice of rotation depends on whether there is a good theoretical reason to suppose that the factors should be related or independent (but see my later comments on this), and also how the variables cluster on the factors before rotation. On the first point, it is probably quite rare that you would measure a set of related variables and expect their underlying dimensions to be completely independent. For example, we wouldn't expect alcoholism to be completely independent of achievement (after all, high achievement leads to high stress, which can lead to the drinks cabinet). Therefore, on theoretical grounds, we should choose oblique rotation. In fact, some argue that oblique rotation is the only sensible choice for naturally occurring data.

On the second point, Figure 17.5 demonstrates how the positioning of clusters is important in determining how successful the rotation will be (note the position of the green circles). If an orthogonal rotation was carried out on the right-hand diagram it would be considerably less successful in maximizing loadings than the oblique rotation that is displayed.

One approach is to run the analysis using both types of rotation. Pedhazur and Schmelkin (1991) suggest that if the oblique rotation demonstrates a negligible correlation between the extracted factors then it is reasonable to use the orthogonally rotated solution. If the oblique rotation reveals a correlated factor structure, then the orthogonally rotated solution should be discarded. We can check the relationships between factors using the **factor transformation matrix**, which is used to convert the unrotated factor loadings into the rotated ones. Values in this matrix represent the angle through which the axes have been rotated, or the degree to which factors have been rotated.

17.4.6.1. Choosing a method of factor rotation ③

SPSS has three methods of orthogonal rotation (**varimax**, **quartimax** and **equamax**) and two methods of oblique rotation (**direct oblimin** and **promax**). These methods differ in how they rotate the factors, so the resulting output depends on which method you select. Quartimax rotation attempts to maximize the spread of factor loadings for a variable across all factors. Therefore, interpreting variables becomes easier. However, this often results in lots of variables loading highly on a single factor. Varimax is the opposite in that it attempts to maximize the dispersion of loadings within factors. Therefore, it tries to load a smaller number of variables highly on each factor, resulting in more interpretable clusters of factors. Equamax is a hybrid of the other two approaches and is reported to behave fairly erratically (see Tabachnick and Fidell, 2012). For a first analysis, you should probably select varimax because it is a good general approach that simplifies the interpretation of factors.

The case with oblique rotations is more complex because correlation between factors is permitted. In the case of direct oblimin, the degree to which factors are allowed to correlate is determined by the value of a constant called delta. The default value in SPSS is 0, and this ensures that high correlation between factors is not allowed (this is known as direct quartimin rotation). If you choose to set delta to greater than 0 (up to 0.8), then you can expect highly correlated factors; if you set delta less than 0 (down to −0.8) you can expect less correlated factors. The default setting of zero is sensible for most analyses, and I don’t recommend changing it unless you know what you are doing (see Pedhazur & Schmelkin, 1991, p. 620). Promax is a faster procedure designed for very large data sets.

In theory, the exact choice of rotation will depend largely on whether or not you think that the underlying factors should be related. If you expect the factors to be independent then you should choose one of the orthogonal rotations (I recommend varimax). If, however, there are theoretical grounds for supposing that your factors might correlate, then direct oblimin should be selected. In practice, there are strong grounds to believe that orthogonal rotations are a complete nonsense for naturalistic data, and certainly for any data involving humans (can you think of any psychological construct that is not in any way correlated with some other psychological construct?) As such, some argue that orthogonal rotations should never be used.

17.4.6.2. Substantive importance of loadings ②

Once a factor structure has been found, it is important to decide which variables make up which factors. Earlier I said that the loadings were a gauge of the substantive importance of a given variable to a given factor. Therefore, it makes sense that we use these values to place variables with factors. It is possible to assess the statistical significance of a loading (after all, it is simply a correlation coefficient or regression coefficient); however, it is not as easy as it seems (see Stevens, 2002, p. 393) because the significance of a factor loading will depend on the sample size. Stevens (2002) produced a table of critical values against which loadings can be compared. To summarize, he recommends that for a sample size of 50 a loading of .722 can be considered significant, for 100 the loading should be greater than .512, for 200 it should be greater than .364, for 300 it should be greater than .298, for 600 it should be greater than .21, and for 1000 it should be greater than .162. These values are based on an alpha level of .01 (two-tailed), which allows for the fact that several loadings will need to be tested (see Stevens, 2002, for further detail). Therefore, in very large samples, small loadings can be considered statistically meaningful.

However, the significance of a loading gives little indication of the substantive importance of a variable to a factor. We can guage importance by squaring the loading to give an estimate of the amount of variance in a factor accounted for by a variable (like R^2). In this respect Stevens (2002) recommends interpreting factor loadings with an absolute value greater than .4 (which explain around 16% of the variance in the variable). Some researchers opt for the lower criterion of .3.

17.5. Research example ②

One of the uses of factor analysis is to develop questionnaires. I have noticed that a lot of students become very stressed about SPSS. Therefore, I wanted to design a questionnaire to measure a trait that I termed ‘SPSS anxiety’. I devised a questionnaire to measure various aspects of students’ anxiety towards learning SPSS, the SAQ (Figure 17.6). I generated questions based on interviews with anxious and non-anxious students and came up with 23 possible questions to include. Each question was a statement followed by a 5-point Likert scale: ‘strongly disagree’, ‘disagree’, ‘neither agree nor disagree’, ‘agree’ and ‘strongly agree’ (SD, D, N, A, and SA, respectively). The questionnaire was designed to measure how anxious a given individual would be about learning how to use SPSS. What’s more, I wanted to know whether anxiety about SPSS could be broken down into specific forms of anxiety. In other words, what latent variables contribute to anxiety about SPSS?

With a little help from a few lecturer friends I collected 2571 completed questionnaires (at this point it should become apparent that this example is fictitious!). Load the data file (**SAQ.sav**) into SPSS and have a look at the variables and their properties. The first thing to note is that each question (variable) is represented by a different column. We know that in SPSS, cases (or people’s data) are stored in rows and variables are stored in columns, so this layout is consistent with past chapters. The second thing to notice is that there are 23 variables labelled **Question_01** to **Question_23** and that each has a label indicating the question. By labelling my variables I can be very clear about what each variable represents (this is the value of giving your variables full titles rather than just using restrictive column headings).

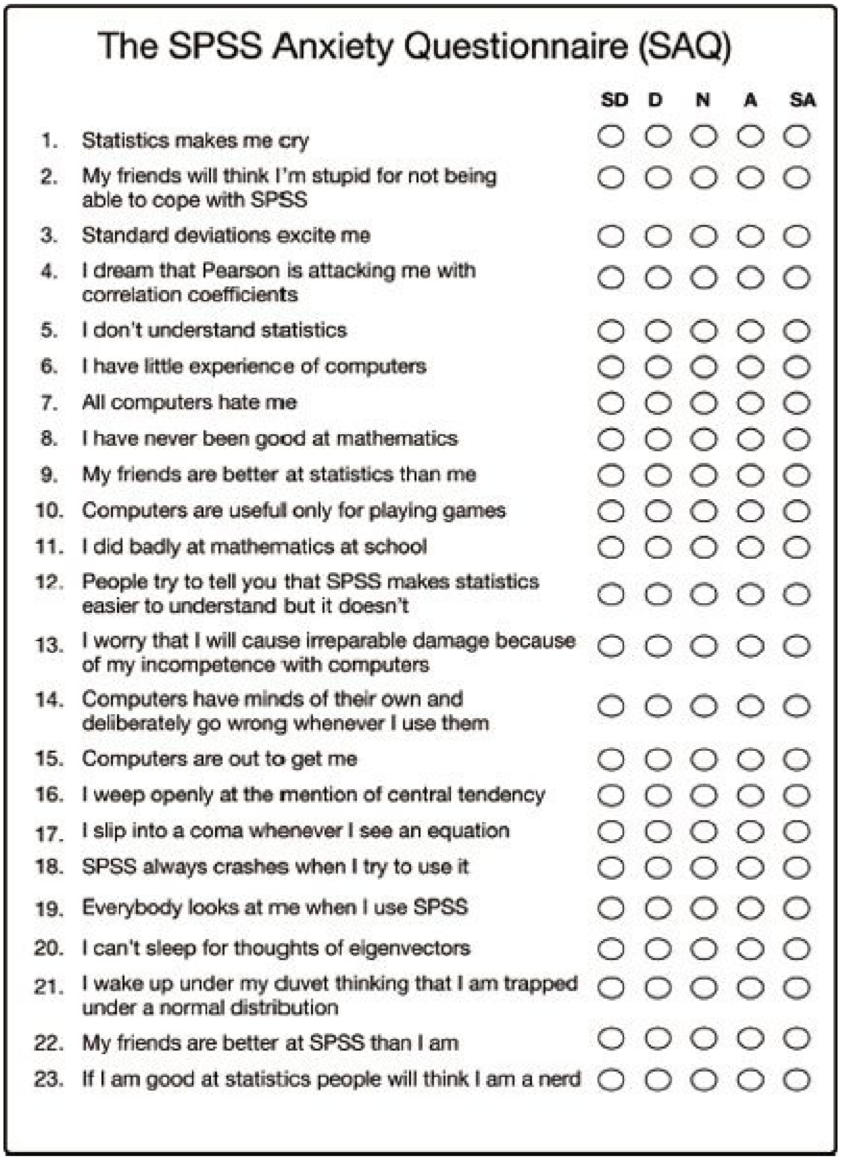
OLIVER TWISTED

Please Sir, can I have some more … questionnaires?

’I’m going to design a questionnaire to measure one’s propensity to pick a pocket or two,’ says Oliver, ‘but how would I go about doing it?’ You’d read the useful information about the dos and don’ts of questionnaire design in the additional material for this chapter on the companion website, that’s how. Rate how useful it is on a Likert scale from 1 = not useful at all, to 5 = very useful.

17.5.1. General procedure ①

Figure 17.7 shows the general procedure for conducting factor analysis or PCA. First we need to do some initial screening of the data, then once we embark on the main analysis we need to consider how many factors to retain and what rotation to use, and if we are using the analysis to look at the factor structure of a questionnaire then we would want to do a reliability analysis at the end (see Section 17.9).



The SPSS Anxiety Questionnaire (SAQ)

		SD	D	N	A	SA
1.	Statistics makes me cry	○	○	○	○	○
2.	My friends will think I'm stupid for not being able to cope with SPSS	○	○	○	○	○
3.	Standard deviations excite me	○	○	○	○	○
4.	I dream that Pearson is attacking me with correlation coefficients	○	○	○	○	○
5.	I don't understand statistics	○	○	○	○	○
6.	I have little experience of computers	○	○	○	○	○
7.	All computers hate me	○	○	○	○	○
8.	I have never been good at mathematics	○	○	○	○	○
9.	My friends are better at statistics than me	○	○	○	○	○
10.	Computers are useful only for playing games	○	○	○	○	○
11.	I did badly at mathematics at school	○	○	○	○	○
12.	People try to tell you that SPSS makes statistics easier to understand but it doesn't	○	○	○	○	○
13.	I worry that I will cause irreparable damage because of my incompetence with computers	○	○	○	○	○
14.	Computers have minds of their own and deliberately go wrong whenever I use them	○	○	○	○	○
15.	Computers are out to get me	○	○	○	○	○
16.	I weep openly at the mention of central tendency	○	○	○	○	○
17.	I slip into a coma whenever I see an equation	○	○	○	○	○
18.	SPSS always crashes when I try to use it	○	○	○	○	○
19.	Everybody looks at me when I use SPSS	○	○	○	○	○
20.	I can't sleep for thoughts of eigenvectors	○	○	○	○	○
21.	I wake up under my duvet thinking that I am trapped under a normal distribution	○	○	○	○	○
22.	My friends are better at SPSS than I am	○	○	○	○	○
23.	If I am good at statistics people will think I am a nerd	○	○	○	○	○

FIGURE 17.6 The SPSS anxiety questionnaire (SAQ)

17.5.2. Before you begin ②

17.5.2.1. Sample size ②

Correlation coefficients fluctuate from sample to sample, much more so in small samples than in large. Therefore, the reliability of factor analysis will depend on sample size. Many 'rules of thumb' exist for the ratio of cases to variables; a common one is to have at least 10–15 participants per variable. Although I've heard this rule bandied about on numerous occasions, its empirical basis is unclear (although Nunnally, 1978, did recommend having 10 times as many participants as variables). Based on real data, Arrindell and van der Ende (1985) concluded that the cases-to-variables ratio made little difference to the stability of factor solutions.

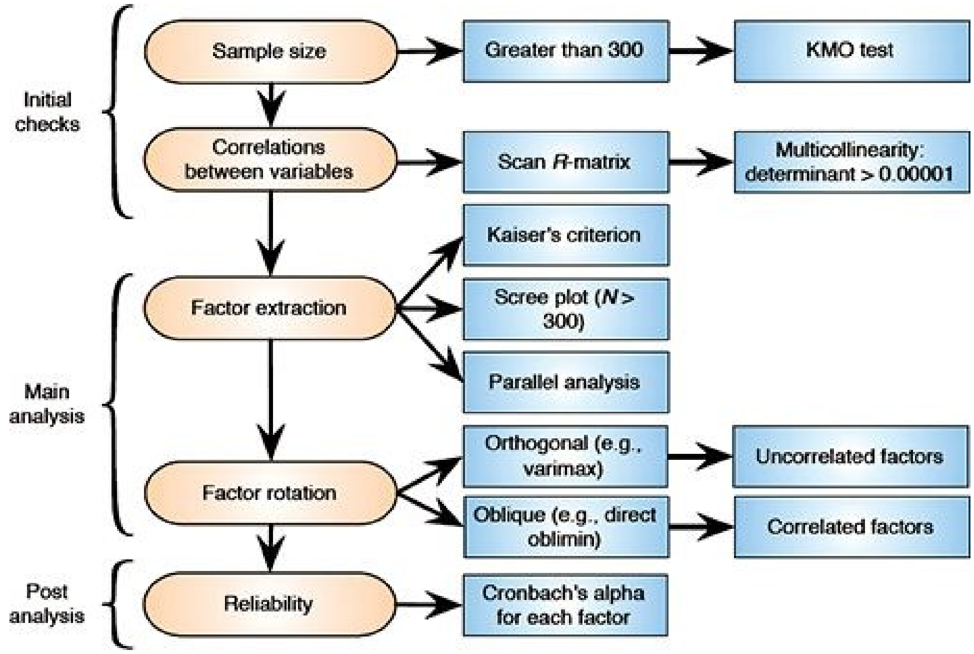


FIGURE 17.7 General procedure for factor analysis and PCA

What does matter is the overall sample size. Test parameters tend to be stable regardless of the cases-to-variables ratio (Kass & Tinsley, 1979), which is why Tabachnick and Fidell (2012) suggest that 'it is comforting to have at least 300 cases' (p. 613) and Comrey and Lee (1992) class 300 as a good sample size, 100 as poor and 1000 as excellent. However, the picture is a little more complicated than that. First, the factor loadings matter: Guadagnoli and Velicer (1988) found that if a factor has four or more loadings greater than .6 then it is reliable regardless of sample size. Furthermore, factors with 10 or more loadings greater than .40 are reliable if the sample size is greater than 150. Finally, factors with a few low loadings should not be interpreted unless the sample size is 300 or more.

Second, the communalities matter. MacCallum, Widaman, Zhang, and Hong (1999) have shown that as communalities become lower the importance of sample size increases. With all communalities above .6, relatively small samples (less than 100) may be perfectly adequate. With communalities in the .5 range, samples between 100 and 200 can be good enough provided there are relatively few factors each with only a small number of indicator variables. In the worst scenario of low communalities (well below .5) and a larger number of underlying factors they recommend samples above 500.

What's clear from this work is that a sample of 300 or more will probably provide a stable factor solution, but that a wise researcher will measure enough variables to measure adequately all of the factors that theoretically they would expect to find.

There are measures of sampling adequacy such as the **Kaiser–Meyer–Olkin measure of sampling adequacy (KMO)** (Kaiser, 1970). The KMO can be calculated for individual and multiple variables and represents the ratio of the squared correlation between variables to the squared partial correlation between variables. The KMO statistic varies between 0 and 1. A value of 0 indicates that the sum of partial correlations is large relative to the sum of correlations, indicating diffusion in the pattern of correlations (hence, factor analysis is likely to be inappropriate). A value close to 1 indicates that patterns of correlations are relatively compact and so factor analysis should yield distinct and reliable factors. Kaiser (1974) recommends accepting values greater than .5 as barely acceptable (values below this should lead you to either collect more data or rethink which variables to include). Hutcheson and Sofroniou (1999) provide appealing guidelines, especially if you like the letter M:

- Marvellous: values in the .90s
- Meritorious: values in the .80s
- Middling: values in the .70s
- Mediocre: values in the .60s
- Miserable: values in the .50s
- Merde: values below .50. (Actually they used the word ‘unacceptable’ but I don’t like the fact that it doesn’t start with the letter ‘M’ so I have changed it.)

17.5.2.2. Correlations between variables ③

When I was an undergraduate, my statistics lecturer always used to say ‘if you put garbage in, you get garbage out’. This saying applies particularly to factor analysis because SPSS will usually find a factor solution to a set of variables. However, the solution is unlikely to have any real meaning if the variables analysed are not sensible. The first thing to do when conducting a factor analysis or PCA is to look at the correlations between variables. There are essentially two potential problems: (1) correlations that are not high enough; and (2) correlations that are too high. In both cases the remedy is to remove variables from the analysis. The correlations between variables can be checked using the *correlate* procedure (see Chapter 7) to create a correlation matrix of all variables. This matrix can also be created as part of the factor analysis. We will look at each problem in turn.

If our test questions measure the same underlying dimension (or dimensions) then we would expect them to correlate with each other (because they are measuring the same thing). Even if questions measure different aspects of the same things (e.g., we could measure overall anxiety in terms of sub-components such as worry, intrusive thoughts and physiological arousal), there should still be high correlations between the variables relating to these sub-traits. We can test for this problem first by visually scanning the correlation matrix and looking for correlations below about .3 (you could use the significance of correlations but, given the large sample sizes normally used with factor analysis, this approach isn’t helpful because even very small correlations will be significant in large samples). If any variables have lots of correlations below .3 then consider excluding them. It should be immediately clear that this approach is very subjective: I’ve used fuzzy terms such as ‘about .3’ and ‘lots of’, but I have to because every data set is different. Analysing data really is a skill, and there’s more to it than following a recipe book!

For an objective test of whether correlations (overall) are too small we can test for a very extreme scenario. If the variables in our correlation matrix did not correlate at all, then our correlation matrix would be an identity matrix (i.e., the off-diagonal components would be zero); so, if the population correlation matrix resembles an identity matrix then it means that every variable correlates very badly with all other variables (i.e., all correlation coefficients are close to zero). **Bartlett’s test** tells us whether our correlation matrix is significantly different from an identity matrix. Therefore, if it is significant then it means that the correlations between variables are (overall) significantly different from zero. The trouble is that because significance depends on sample size (see Section 2.6.1.10) and in factor analysis sample sizes are very large, Bartlett’s test will nearly always be significant: even when the correlations between variables are very small indeed. As such, it’s not a useful test (although in the unlikely event that it is non-significant then you certainly have a big problem).

The opposite problem is when variables correlate too highly. Although mild multicollinearity is not a problem for factor analysis it is important to avoid extreme multicollinearity (i.e., variables that are very highly correlated) and **singularity** (variables that are perfectly correlated). As with regression, multicollinearity causes problems in factor analysis because it becomes impossible to determine the unique contribution to a factor of the variables that are highly correlated. Multicollinearity does not cause a problem for PCA.

Multicollinearity can be detected by looking at the determinant of the *R*-matrix, denoted *R* (see Jane Superbrain Box 17.3). One simple heuristic is that the determinant of the *R*-matrix should be greater than 0.00001.

To try to avoid or to correct for multicollinearity you could look through the correlation matrix for variables that correlate very highly ($r > .8$) and consider eliminating one of the variables (or more depending on the extent of the problem) before proceeding. The problem with a heuristic such as this is that the effect of two variables correlating with $r = .9$ might be less than the effect of, say, three variables that all correlate at $r = .6$. In other words, eliminating such highly correlating variables might not be getting at the cause of the multicollinearity (Rockwell, 1975). It may take trial and error to work out which variables are creating the problem.

17.5.2.3. The distribution of data ②

As well as looking for interrelations, you might ensure that variables have roughly normal distributions and are measured at an interval level (which Likert scales are, perhaps wrongly, assumed to be). The assumption of normality is important if you wish to generalize the results of your analysis beyond the sample collected or do significance tests, but otherwise it's not. You can do factor analysis on non-continuous data; for example, if you had dichotomous variables, it's possible (using syntax) to do the factor analysis direct from the correlation matrix, but you should construct the correlation matrix from tetrachoric correlation coefficients (http://www.john-uebersax.com/stat/tetra.htm). The only hassle is computing the correlations (but see the website for software options).

17.6. Running the analysis ②

Access the main dialog box (Figure 17.9) by selecting Analyze Dimension Reduction ▸ Factor... Simply select the variables you want to include in the analysis (remember to exclude any variables that were identified as problematic during the data screening) and transfer them to the box labelled *Variables* by clicking on .

There are several options available, the first of which can be accessed by clicking on Descriptives... to access the dialog box in Figure 17.10. The *Univariate descriptives* option provides means and standard deviations for each variable. Most of the other options relate to the correlation matrix of variables (the *R*-matrix described earlier). The *Coefficients* option produces the *R*-matrix, and selecting the *Significance levels* option will include the significance value of each correlation in the *R*-matrix. You can also ask for the *Determinant* of this matrix, which is useful for testing for multicollinearity or singularity (see Section 17.5.2.2).

KMO and Bartlett's test of sphericity produces the Kaiser–Meyer–Olkin (see Section 17.5.2.1) measure of sampling adequacy and Bartlett's test (see Section 17.5.2.2). We have already seen the various criteria for adequacy, but with a sample of 2571 we shouldn't have cause to worry.

The *Reproduced* option produces a correlation matrix based on the model (rather than the real data). Differences between the matrix based on the model and the matrix based on the observed data indicate the residuals of the model. SPSS produces these residuals in the lower table of the reproduced matrix, and we want relatively few of these values to be greater than .05. Luckily, to save us scanning this matrix, SPSS produces a summary of how many residuals lie above .05. The *Reproduced* option should be selected to obtain this summary. The *Anti-image* option produces an anti-image matrix of covariances and correlations. These matrices contain measures of sampling adequacy for each variable along the diagonal and the negatives of the partial correlation/covariances on the off-diagonals. The diagonal elements, like the KMO measure, should all be greater than .5 at a bare minimum if the sample is adequate for a given pair of variables. If any pair of variables has a value less than this, consider dropping one of them from the analysis. The off-diagonal elements should all be very small (close to zero) in a good model. When you have finished with this dialog box click on Continue to return to the main dialog box.

JANE SUPERBRAIN 17.3

What is the determinant? ③

The determinant of a matrix is an important diagnostic tool in factor analysis, but the question of what it is is not easy to answer because it has a mathematical definition and I'm not a mathematician. However, we can bypass the maths and think about the determinant conceptually. The way that I think of the determinant is as describing the 'area' of the data. In Jane Superbrain Box 8.3 we saw the two diagrams in Figure 17.8. At the time I used these to describe eigenvectors and eigenvalues (which describe the shape of the data). The determinant is related to eigenvalues and eigenvectors but instead of describing the height and width of the data it describes the overall area. So, in the left diagram, the determinant of those data would represent the area inside the red dashed ellipse. These variables have a low correlation so the determinant (area) is big; the biggest value it can be is 1. In the right diagram, the variables are perfectly correlated or singular, and the ellipse (red dashed line) has been squashed down to basically a straight line. In other words, the opposite sides of the ellipse have actually met each other and there is no distance between them at all. Put another way, the area, or determinant, is zero. Therefore, the determinant tells us whether the correlation matrix is singular (determinant is 0), or if all variables are completely unrelated (determinant is 1), or somewhere in between.

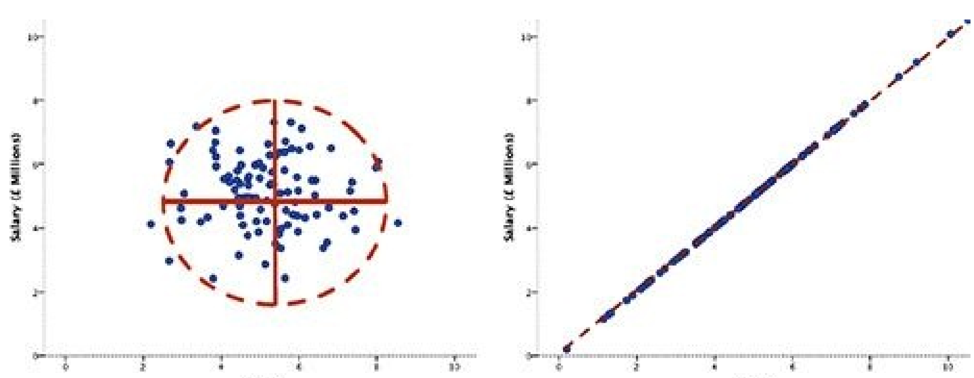


FIGURE 17.8 Data with a large (left) and small (right) determinant

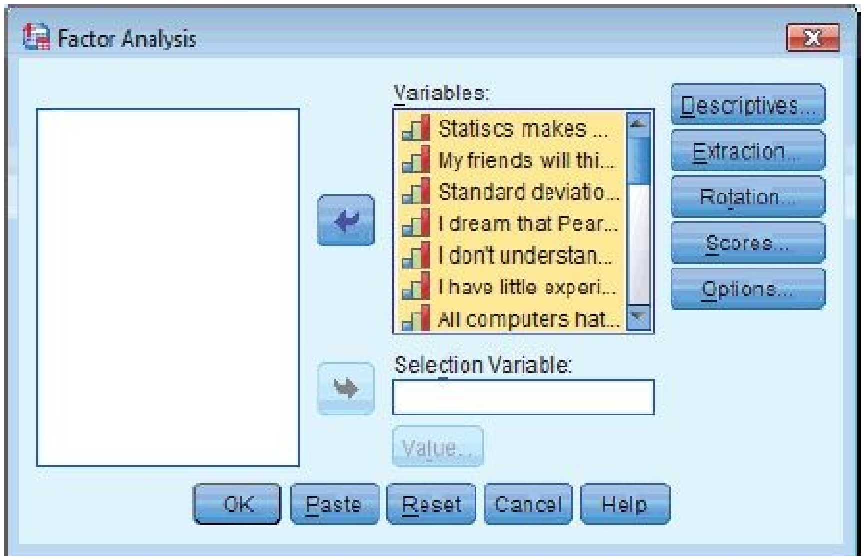


FIGURE 17.9
Main dialog box for factor analysis

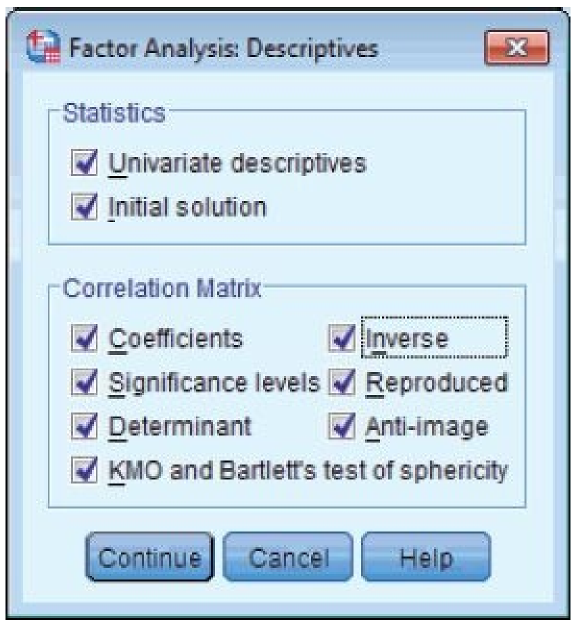


FIGURE 17.10
Descriptives in factor analysis

17.6.1. Factor extraction in SPSS ②

To access the *Extraction* dialog box (Figure 17.11), click on Extraction... in the main dialog box. There are several ways of conducting a factor analysis (see Section 17.4.1). For our purposes we will use *principal axis factoring* (Principal axis factoring). In the *Analyze* box there are two options: to analyse the *Correlation matrix* or to analyse the *Covariance matrix* (SPSS Tip 17.1). The *Display* box has two options within it: to display the *Unrotated factor solution* and a *Scree plot*. The scree plot was described in Section 17.4.5 and is a useful way of establishing how many factors should be retained in an analysis. The factor solution is useful in assessing the improvement of interpretation due to rotation. If the rotated solution is little better than the unrotated solution then it is possible that an inappropriate (or less optimal) rotation method has been used.

SPSS TIP 17.1 Correlation or covariance matrix? ③

You should be happy with the idea that the variance–covariance matrix and correlation matrix are different versions of the same thing. However, generally the results will differ depending on which matrix you analyse. Analysing the correlation matrix is a useful default method because it takes the standardized form of the matrix; therefore, if variables have been measured using different scales this will not affect the analysis. In this example, all variables have been measured using the same measurement scale (a 5-point Likert scale), but often you will want to analyse variables that use different measurement scales. Analysing the correlation matrix ensures that differences in measurement scales are accounted for. In addition, even variables measured using the same scale can have very different variances and this creates problems for PCA. Using the correlation matrix eliminates this problem also.

Having said that, there are statistical reasons for preferring to analyse the covariance matrix: correlation coefficients are not sensitive to variations in the dispersion of data, whereas the covariance is and so it produces better-defined factor structures (Tinsley & Tinsley, 1987). However, the covariance matrix should be analysed only when your variables are commensurable.

The *Extract* box provides options pertaining to the retention of factors. You have the choice of either selecting factors with eigenvalues greater than a user-specified value or retaining a fixed number of factors. For the *Eigenvalues greater than* option the default is Kaiser's recommendation of eigenvalues over 1, but you could change this to Jolliffe's recommendation of 0.7 or any other value you want. It is probably best to run a primary analysis with the *Eigenvalues greater than* 1 option selected, select a scree plot and compare the results. If looking at the scree plot and the eigenvalues over 1 lead you to retain the same number of factors then continue with the analysis and be happy. If the two criteria give different results then examine the communalities and decide for yourself which of the two criteria to believe. If you decide to use the scree plot then you may need to redo the analysis specifying the number of factors to extract. The number of factors to be extracted can be specified by selecting *Fixed number of factors* and then typing the appropriate number in the space provided (e.g., 4).

17.6.2. Rotation ②

We have already seen that the interpretability of factors can be improved through rotation (Section 17.4.6). Click on Rotation... to access the dialog box in Figure 17.12. I've discussed the various rotation options in Section 17.4.6.1, but, to summarize, if there are theoretical grounds to think that the factors are independent (unrelated) then you should choose one of the orthogonal rotations (I recommend varimax) but if theory suggests that your factors might correlate then one of the oblique rotations (direct oblimin or promax) should be selected. In this example I've selected varimax.

The dialog box also has options for displaying the *Rotated solution* and a *Loading plot*. The rotated solution is displayed by default and is essential for interpreting the final rotated analysis. The loading plot will provide a graphical display of each variable plotted against the extracted factors up to a maximum of three factors (unfortunately SPSS cannot produce four- or five-dimensional graphs). This plot is basically similar to Figure 17.3 and it uses the factor loading of each variable for each factor.

With two factors these plots are fairly interpretable, and you should hope to see one group of variables clustered close to the *X*-axis and a different group of variables clustered around the *Y*-axis. If all variables are clustered between the axes, then the rotation has been relatively unsuccessful in maximizing the loading of a variable on a single factor. With three factors these plots will strain even the most dedicated visual system, so unless you have only two factors I would probably avoid them.

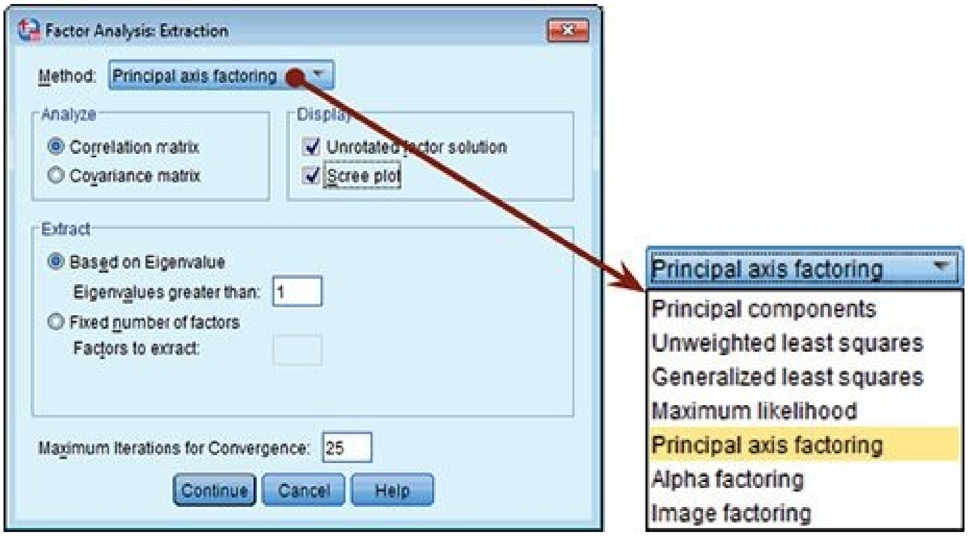


FIGURE 17.11
Dialog box for factor extraction

A final option is to set the *Maximum Iterations for Convergence* (see SPSS Tip 19.1), which specifies the number of times that the computer will search for an optimal solution. In most circumstances the default of 25 is adequate; however, if you get an error message about convergence then increase this value.

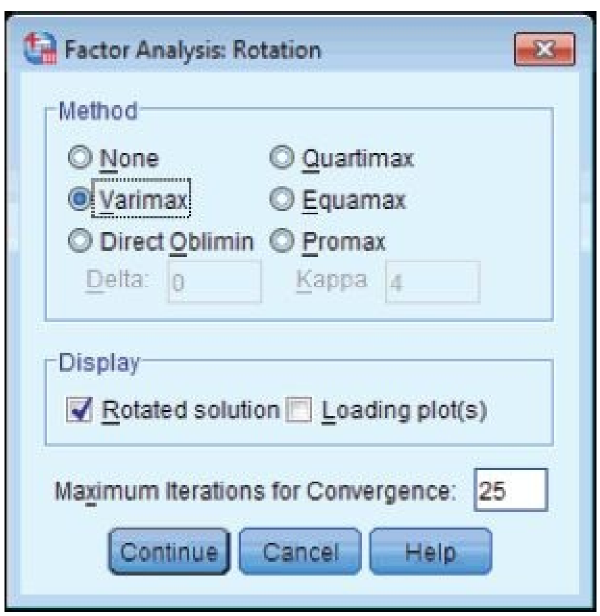


FIGURE 17.12
Factor Analysis: Rotation dialog box

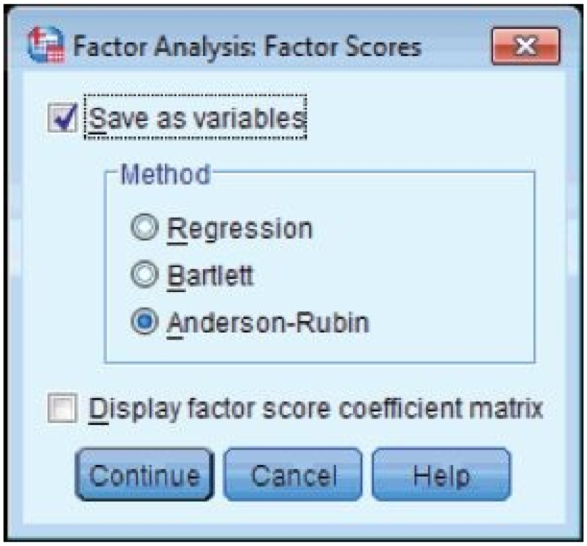


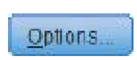
FIGURE 17.13
Factor Analysis: Factor Scores dialog box

17.6.3. Scores ②

The *Factor Scores* dialog box (Figure 17.13) can be accessed by clicking on 

in the main dialog box. This option allows you to save factor scores (see Section 17.3.3) for each case in the data editor. SPSS creates a new column for each factor extracted and then places the factor score for each case within that column. These scores can then be used for further analysis, or simply to identify groups of participants who score highly on particular factors. There are three methods of obtaining these scores, all of which were described in Section 17.3.3. If you want to ensure that factor scores are uncorrelated then select the *Anderson-Rubin* method; if correlations between factor scores are acceptable then choose the *Regression* method. As a final option, you can ask SPSS to produce the factor score coefficient matrix. This matrix is used to compute the factor scores, but realistically, we don't need to see it.

17.6.4. Options ②

The *Options* dialog box can be obtained by clicking on 

in the main dialog box (Figure 17.14). Missing data are a problem for factor analysis just like most other procedures, and SPSS provides a choice of excluding cases or estimating a value for a case. Tabachnick and Fidell (2012) have an excellent chapter on data screening (see also the rather less excellent Chapter 5 of this book). Based on their advice, you should consider the distribution of missing data. If the missing data are non-normally distributed or the sample size after exclusion is too small then estimation is necessary. SPSS uses the mean as an estimate (*Replace with mean*). These procedures lower the standard deviation of variables and so can lead to significant results that would otherwise be non-significant. Therefore, if missing data are random, you might consider excluding cases. SPSS allows you to either *Exclude cases listwise*, in which case any participant with missing data for any variable is excluded, or to *Exclude cases pairwise*, in which case a participant's data are excluded only from calculations for which a datum is missing (see SPSS Tip 5.1). If you exclude cases pairwise your estimates can go all over the place, so it's probably safest to opt to exclude cases listwise unless this results in a massive loss of data.

The final two options relate to how coefficients are displayed. By default, SPSS will list variables in the order in which they are entered into the data editor. However, when interpreting factors it is useful to list variables by size. By selecting *Sorted by size*, SPSS will order the variables by their factor loadings. In fact, it does this sorting fairly intelligently so that all of the variables that load highly on the same factor are displayed together. The second option is to *Suppress absolute values less than* a specified value (by default 0.1). This option ensures that factor loadings within ±0.1 are not displayed in the output. Again, this option is useful for interpretation. The default value is probably sensible, but on your first analysis I recommend changing it either to .3 or to a value reflecting the expected value of a significant factor loading given the sample size (see Section 17.4.6.2). This will make interpretation simpler. We know that a loading of .4 is substantial, but so we don't throw out the baby with the bath water, setting the value to 0.3 is sensible: we will see not only the substantial loadings but those close to the cut-off (e.g., a loading of .39). For this example set the value at .3.

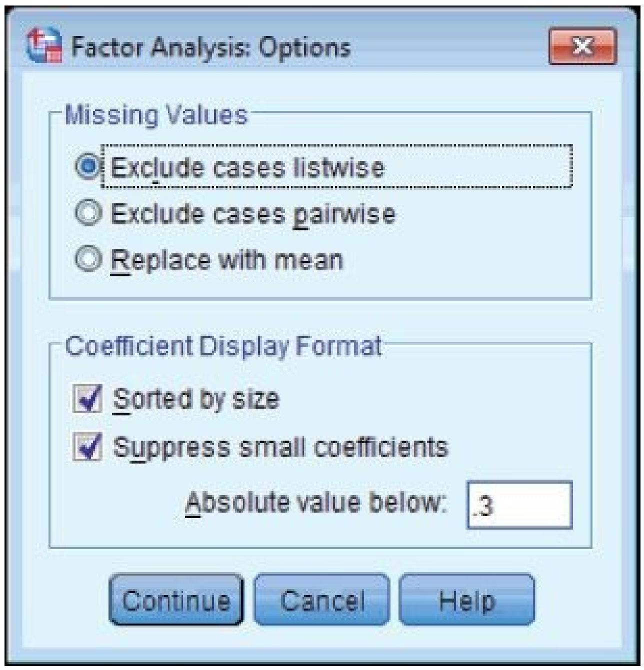


FIGURE 17.14
Factor Analysis: options dialog box

ODITI'S LANTERN

PCA

'I, Oditi, feel that we are getting closer to finding the hidden truths behind the numbers. Factor analysis allows us to estimate variables "hidden" within the data. This technique is the very essence of the cult of undiscovered numerical truths. Once we have mastered this tool we can find out what people are really thinking, even if they don't know they're thinking it. We might find that they think that they think I'm the kind saviour of cute furry gerbils, but that underneath they know the truth … stare into my lantern to discover factor analysis.'

17.7. Interpreting output from SPSS ②

Select the same options as I have in the screen diagrams and run a factor analysis with orthogonal rotation.

SELF-TEST Having done this, select the *Direct Oblimin* option in Figure 17.12 and repeat the analysis. You should obtain two outputs identical in all respects except that one used an orthogonal rotation and the other an oblique.

To save space I set the default SPSS options such that each variable is referred to only by its label on the data editor (e.g., Question_12). On the output *you* obtain, you should find that the SPSS uses the value label (the question itself) in all of the output. When using the output refer back to Figure 17.6 to remind you of what each question was.

When you factor-analyse your own data, you might be unlucky enough to see an error message about a 'non-positive definite matrix' (see SPSS Tip 17.2). A 'non-positive definite matrix' sounds a bit like a collection of depressed numbers that lack certainty about their lives. In some ways it is.

17.7.1. Preliminary analysis ②

The first body of output concerns data screening, assumption testing and sampling adequacy. You'll find several large tables (or matrices) that tell us interesting things about our data. If you selected the *Univariate descriptives* option in Figure 17.10 then the first table will contain descriptive statistics for each variable (the mean, standard deviation and number of cases). This table is not included here, but you should have enough experience to be able to interpret it. The table also includes the number of missing cases; this summary is a useful way to determine the extent of missing data.

SPSS TIP 17.2 Error messages about a 'non-positive definite matrix' ④

Factor analysis works by looking at your correlation matrix. This matrix has to be 'positive definite' for the analysis to work. This term means lots of horrible things mathematically (e.g., the eigenvalues and determinant of the matrix have to be positive), but, in more basic terms, factors are like lines floating in space, and eigenvalues measure the length of those lines. If your eigenvalue is negative then it means that the length of your line/factor is negative too. It's a bit like me asking you how tall you are, and you responding 'I'm minus 175 cm tall'. That would be nonsense. If a factor has negative length, then that is nonsense too. When SPSS decomposes the correlation matrix to look for factors, if it comes across a negative eigenvalue it starts thinking 'oh dear, I've entered some weird parallel universe where the usual rules of maths no longer apply and things can have negative lengths, and this probably means that time runs backwards, my mum is my dad, my sister is a dog, my head is a fish, and my toe is a frog called Gerald'. It does the sensible thing and decides not to proceed. Things like the KMO test and the determinant rely on a positive definite matrix; if you don't have one they can't be computed.

The most likely reason for having a non-positive definite *R*-matrix is that you have too many variables and too few cases of data, which makes the correlation matrix a bit unstable. It could also be that you have too many highly correlated items in your matrix (singularity, for example, tends to mess things up). In any case it means that your data are bad, naughty data, and not to be trusted; if you let them loose then you have only yourself to blame for the consequences.

Other than cry, there's not that much you can do to rectify the situation. You could try to limit your items, or selectively remove items (especially highly correlated ones) to see if that helps. Collecting more data can help too. There are some mathematical fudges you can do, but they're not as tasty as vanilla fudge and they are hard to implement.

Correlation Matrix[a]

		Question_01	Question_02	Question_03	Question_04	Question_05	Question_19	Question_20	Question_21	Question_22	Question_23
Correlation	Question_01	1.000	-.099	-.337	.436	.402	-.189	.214	.329	-.104	-.004
	Question_02	-.099	1.000	.318	-.112	-.119	.203	-.202	-.205	.231	.100
	Question_03	-.337	.318	1.000	-.380	-.310	.342	-.325	-.417	.204	.150
	Question_04	.436	-.112	-.380	1.000	.401	-.186	.243	.410	-.098	-.034
	Question_05	.402	-.119	-.310	.401	1.000	-.165	.200	.335	-.133	-.042
	Question_06	.217	-.074	-.227	.278	.257	-.167	.101	.272	-.165	-.069
	Question_07	.305	-.159	-.382	.409	.339	-.269	.221	.483	-.168	-.070
	Question_08	.331	-.050	-.259	.349	.269	-.159	.175	.296	-.079	-.050
	Question_09	-.092	.315	.300	-.125	-.096	.249	-.159	-.136	.257	.171
	Question_10	.214	-.084	-.193	.216	.258	-.127	.084	.193	-.131	-.062
	Question_11	.357	-.144	-.351	.369	.298	-.200	.255	.346	-.162	-.086
	Question_12	.345	-.195	-.410	.442	.347	-.267	.298	.441	-.167	-.046
	Question_13	.355	-.143	-.318	.344	.302	-.227	.204	.374	-.195	-.053
	Question_14	.338	-.165	-.371	.351	.315	-.254	.226	.399	-.170	-.048
	Question_15	.246	-.165	-.312	.334	.261	-.210	.206	.300	-.168	-.062
	Question_16	.499	-.168	-.419	.416	.395	-.267	.265	.421	-.156	-.082
	Question_17	.371	-.087	-.327	.383	.310	-.163	.205	.363	-.126	-.092
	Question_18	.347	-.164	-.375	.382	.322	-.257	.235	.430	-.160	-.080
	Question_19	-.189	.203	.342	-.186	-.165	1.000	-.249	-.275	.234	.122
	Question_20	.214	-.202	-.325	.243	.200	-.249	1.000	.468	-.100	-.035
	Question_21	.329	-.205	-.417	.410	.335	-.275	.468	1.000	-.129	-.068
	Question_22	-.104	.231	.204	-.098	-.133	.234	-.100	-.129	1.000	.230
	Question_23	-.004	.100	.150	-.034	-.042	.122	-.035	-.068	.230	1.000
Sig. (1-tailed)	Question_01		.000	.000	.000	.000	.000	.000	.000	.000	.410
	Question_02	.000		.000	.000	.000	.000	.000	.000	.000	.000
	Question_03	.000	.000		.000	.000	.000	.000	.000	.000	.000
	Question_04	.000	.000	.000		.000	.000	.000	.000	.000	.043
	Question_05	.000	.000	.000	.000		.000	.000	.000	.000	.017
	Question_06	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Question_07	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Question_08	.000	.006	.000	.000	.000	.000	.000	.000	.000	.005
	Question_09	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Question_10	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001
	Question_11	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Question_12	.000	.000	.000	.000	.000	.000	.000	.000	.000	.009
	Question_13	.000	.000	.000	.000	.000	.000	.000	.000	.000	.004
	Question_14	.000	.000	.000	.000	.000	.000	.000	.000	.000	.007
	Question_15	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001
	Question_16	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Question_17	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Question_18	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Question_19	.000	.000	.000	.000	.000		.000	.000	.000	.000
	Question_20	.000	.000	.000	.000	.000	.000		.000	.000	.039
	Question_21	.000	.000	.000	.000	.000	.000	.000		.000	.000
	Question_22	.000	.000	.000	.000	.000	.000	.000	.000		.000
	Question_23	.410	.000	.000	.043	.017	.000	.039	.000	.000	

a. Determinant = .001

OUTPUT 17.1

Output 17.1 shows the *R*-matrix (i.e., the correlation matrix)[6] produced using the *Coefficients* and *Significance levels* options in Figure 17.10. The top half of this table contains the Pearson correlation coefficient between all pairs of questions, whereas the bottom half contains the one-tailed significance of these coefficients. We can use this correlation matrix to check the pattern of relationships. First, scan the matrix for correlations greater than .3, and look for variables that only have a small number of correlations greater than this value. Then scan the correlation coefficients themselves and look for any greater than .9. If any are found then you should be aware that a problem could arise because of multicollinearity in the data.

You can also check the determinant of the correlation matrix and, if necessary, eliminate variables that you think are causing the problem. The determinant is listed at the bottom of the matrix (blink and you'll miss it). For these data its value is .001, which is greater than the necessary value of 0.00001 (see Section 17.6).[7] To sum up, all questions in the SAQ correlate reasonably well with all others and none of the correlation coefficients are excessively large; therefore, we won't eliminate any questions at this stage.

If you selected the *Inverse* option in Figure 17.10 you'll find the inverse of the correlation matrix (R^{-1}) in your output (labelled *Inverse of Correlation Matrix*). This matrix is used in various calculations (including factor scores – see Section 17.3.3.1), but in all honesty is useful only if you want some insight into the calculations that go on in a factor analysis. Most of us have more interesting things to do, so ignore it.

If you selected the *KMO and Bartlett's test of sphericity* and the *Anti-image* options in Figure 17.10 then your output will contain the Kaiser–Meyer–Olkin measure of sampling adequacy and Bartlett's test of sphericity (Output 17.2) and the anti-image correlation and covariance matrices (an edited version is in Output 17.3). The anti-image correlation and covariance matrices provide similar information (remember the relationship between covariance and correlation) and so only the anti-image correlation matrix need be studied in detail because it is the most informative.

For the KMO statistic the value is .93, which is well above the minimum criterion of .5 and falls into the range of 'marvellous' (see Section 17.5.2.1), so we should be confident that the sample size is adequate for factor analysis. I mentioned before that KMO can be calculated for multiple and individual variables. The KMO values for individual variables are produced on the diagonal of the anti-image correlation matrix (I have highlighted these cells in Output 17.3). As well as checking the overall KMO statistic, we should examine the diagonal elements of the anti-image correlation matrix: the values should all be above the bare minimum of .5 (and preferably higher). For these data all values are well above .5, which is good news. If you find any variables with values below 0.5 then you should consider excluding them from the analysis (or run the analysis with and without that variable and note the difference). Removal of a variable affects the KMO statistics, so if you do remove a variable be sure to re-examine the new anti-image correlation matrix. As for the rest of the anti-image correlation matrix, the off-diagonal elements represent the partial correlations between variables. For a good factor analysis we want these correlations to be very small (the smaller, the better). So, as a final check you can look through to see that the off-diagonal elements are small (they should be for these data).

Bartlett's measure (Output 17.2) tests the null hypothesis that the original correlation matrix is an identity matrix. We want this test to be *significant* (see Section 17.5.2.2). As I mentioned before, given the large sample sizes usually used in factor analysis this test will almost certainly be significant, and it is ($p < .001$). A non-significant test would certainly indicate a massive problem, but this significant value only really tells us that we don't have a massive problem, which is nice to know, I suppose.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.930
Bartlett's Test of Sphericity	Approx. Chi-Square	19334.492
	df	253
	Sig.	.000

OUTPUT 17.2

Anti-image Matrices

Anti-image Correlation

	Question_01	Question_02	Question_03	Question_04	Question_05	Question_19	Question_20	Question_21	Question_22	Question_23
Question_01	.930	-.020	.053	-.167	-.156	.012	-.016	.006	.001	-.059
Question_02	-.020	.875	-.157	-.041	.010	-.029	.059	.041	-.121	-.002
Question_03	.053	-.157	.951	.084	.037	-.121	.078	.070	-.007	-.076
Question_04	-.167	-.041	.084	.955	-.134	-.034	-.004	-.086	-.033	-.017
Question_05	-.156	.010	.037	-.134	.960	-.018	-.011	-.046	.035	-.005
Question_06	.020	-.053	-.042	-.007	-.035	-.015	.051	.019	.040	.018
Question_07	.023	.016	.072	-.087	-.044	.068	.048	-.208	.013	-.008
Question_08	-.049	-.033	-.007	-.075	-.027	.047	.021	-.020	-.023	.002
Question_09	-.016	-.193	-.142	.030	-.020	-.111	.038	-.031	-.126	-.092
Question_10	-.012	-.012	-.016	.006	-.093	-.009	.043	.017	.019	.015
Question_11	-.041	.038	.064	-.022	.000	-.006	-.082	-.005	.034	.010
Question_12	-.007	.031	.087	-.154	-.058	.040	-.065	-.079	.018	-.028
Question_13	-.085	-.008	-.032	.023	.004	.009	.018	-.033	.052	-.030
Question_14	-.040	.023	.069	-.004	-.026	.044	.001	-.063	.029	-.026
Question_15	.089	.037	.008	-.062	.014	.009	-.037	.035	.025	-.024
Question_16	-.264	-.011	.081	-.036	-.095	.047	-.005	-.085	-.003	.023
Question_17	-.047	-.029	.015	-.035	-.018	-.047	.015	-.041	.010	.055
Question_18	-.023	.018	.039	-.025	.002	.030	-.003	-.072	-.024	.023
Question_19	.012	-.029	-.121	-.034	-.018	.941	.091	.031	-.115	-.038
Question_20	-.016	.059	.078	-.004	-.011	.091	.889	-.323	-.011	-.028
Question_21	.006	.041	.070	-.086	-.046	.031	-.323	.929	-.024	.013
Question_22	.001	-.121	-.007	-.033	.035	-.115	-.011	-.024	.878	-.176
Question_23	-.059	-.002	-.076	-.017	-.005	-.038	-.028	.013	-.176	.766

OUTPUT 17.3

CRAMMING SAM'S TIPS Preliminary analysis

- Scan the correlation matrix; look for variables that don't correlate with any other variables, or correlate very highly (r = .9) with one or more other variables.
- In factor analysis, check that the determinant of this matrix is bigger than 0.00001; if it is then multicollinearity isn't a problem.
- In the table labelled *KMO and Bartlett's Test* the KMO statistic should be greater than .5 as a bare minimum; if it isn't, collect more data. You should check the KMO statistic for individual variables by looking at the diagonal of the anti-image matrices again, these values should be above .5 (this is useful for identifying problematic variables if the overall KMO is unsatisfactory).
- Bartlett's test of sphericity will usually be significant (the value of *Sig.* will be less than .05); if it's not you've got a disaster on your hands.

17.7.2. Factor extraction ②

The first part of the factor extraction process is to determine the linear components within the data set (the eigenvectors) by calculating the eigenvalues of the *R*-matrix (see Section 17.4.4). We know that there are as many components (eigenvectors) in the *R*-matrix as there are variables, but most will be unimportant. To determine the importance of a particular vector we look at the magnitude of the associated eigenvalue. We can then apply criteria to determine which factors to retain and which to discard. By default SPSS uses Kaiser's criterion of retaining factors with eigenvalues greater than 1 (see Figure 17.11).

Output 17.4 lists the eigenvalues associated with each factor before extraction, after extraction and after rotation. Before extraction, SPSS has identified 23 factors within the data set (we know that there should be as many eigenvectors as there are variables and so there will be as many factors as variables – see Section 17.4.4). The eigenvalues associated with each factor represent the variance explained by that particular factor; SPSS also displays the eigenvalue in terms of the percentage of variance explained (so factor 1 explains 31.696% of total variance). The first few factors explain relatively large amounts of variance (especially factor 1), whereas subsequent factors explain only small amounts of variance. SPSS then extracts all factors with eigenvalues greater than 1, which leaves us with four factors. The eigenvalues associated with these factors are again displayed (and the percentage of variance explained) in the columns labelled *Extraction Sums of Squared Loadings*. In the final part of the table (labelled *Rotation Sums of Squared Loadings*), the eigenvalues of the factors after rotation are displayed. Rotation has the effect of optimizing the factor structure, and one consequence for these data is that the relative importance of the four factors is equalized a bit. Before rotation, factor 1 accounted for considerably more variance than the remaining three (29.32% compared to 4.90%, 3.54% and 2.71%), but after rotation it accounts for only 13.19% of variance (compared to 12.42%, 8.64% and 6.24%, respectively).

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.290	31.696	31.696	6.744	29.323	29.323	3.033	13.188	13.188
2	1.739	7.560	39.256	1.128	4.902	34.225	2.855	12.415	25.603
3	1.317	5.725	44.981	.814	3.539	37.764	1.986	8.636	34.238
4	1.227	5.336	50.317	.624	2.713	40.477	1.435	6.239	40.477
5	.988	4.295	54.612						
6	.895	3.893	58.504						
7	.806	3.502	62.007						
8	.783	3.404	65.410						
9	.751	3.265	68.676						
10	.717	3.117	71.793						
11	.684	2.972	74.765						
12	.670	2.911	77.676						
13	.612	2.661	80.337						
14	.578	2.512	82.849						
15	.549	2.388	85.236						
16	.523	2.275	87.511						
17	.508	2.210	89.721						
18	.456	1.982	91.704						
19	.424	1.843	93.546						
20	.408	1.773	95.319						
21	.379	1.650	96.969						
22	.364	1.583	98.552						
23	.333	1.448	100.000						

Extraction Method: Principal Axis Factoring.

OUTPUT 17.4

Output 17.5 (left) shows the table of communalities before and after extraction. Remember that the communality is the proportion of common variance within a variable (see Section 17.4.1). Factor analysis starts by estimating the variance that is common; therefore, before extraction the communalities are a kind of best guess. Once factors have been extracted, we have a better idea of how much variance is, in reality, common. The communalities in the column labelled *Extraction* reflect this common variance. So, for example, we can say that 37.3% of the variance associated with question 1 is common, or shared, variance. Another way to look at these communalities is in terms of the proportion of variance explained by the underlying factors. Remember that after extraction we have discarded some factors (in this case we've retained only four), so the communalities after extraction represent the amount of variance in each variable that can be explained by the retained factors.

Communalities

	Initial	Extraction
Question_01	.373	.373
Question_02	.188	.260
Question_03	.398	.472
Question_04	.385	.419
Question_05	.291	.299
Question_06	.427	.594
Question_07	.470	.489
Question_08	.490	.646
Question_09	.220	.339
Question_10	.197	.197
Question_11	.530	.629
Question_12	.424	.453
Question_13	.451	.474
Question_14	.393	.425
Question_15	.344	.322
Question_16	.463	.458
Question_17	.494	.575
Question_18	.492	.544
Question_19	.209	.245
Question_20	.270	.266
Question_21	.454	.468
Question_22	.167	.247
Question_23	.086	.116

Extraction Method: Principal Axis Factoring.

Factor Matrix[a]

	Factor			
	1	2	3	4
Question_18	.684			
Question_07	.663			
Question_16	.653			
Question_13	.650			
Question_11	.646	.313		
Question_12	.643			
Question_21	.633			
Question_17	.632	.359		
Question_14	.628			
Question_04	.607			
Question_03	-.605			
Question_15	.559			
Question_01	.557			
Question_06	.552		.489	
Question_08	.546	.483		
Question_05	.522			
Question_20	.407			
Question_10	.404			
Question_19	-.397			
Question_09		.460		
Question_02		.372		
Question_22				
Question_23				

Extraction Method: Principal Axis Factoring.
a. 4 factors extracted. 11 iterations required.

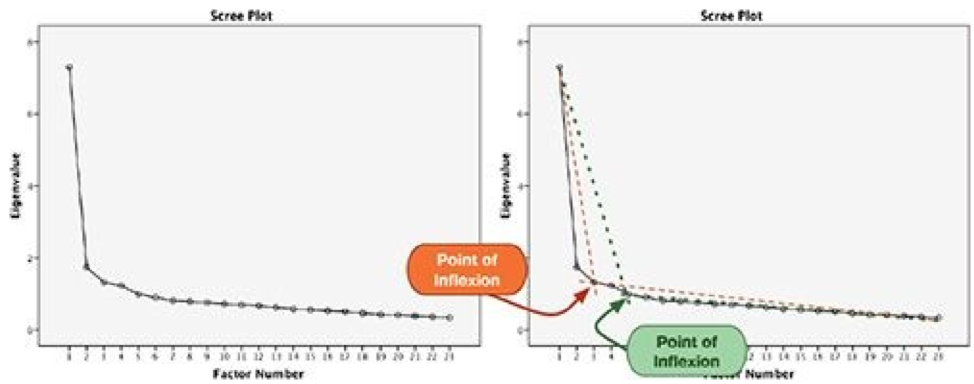
OUTPUT 17.5

Output 17.5 (right) also shows the factor matrix before rotation. This matrix contains the loadings of each variable on each factor. By default SPSS displays all loadings; however, we requested that all loadings less than .3 be suppressed in the output (see Figure 17.14) and so there are blank spaces for many of the loadings. This matrix is not particularly important for interpretation, but it is interesting to note that before rotation most variables load highly on the first factor (that is why this factor accounts for most of the variance in Output 17.4).

Factor analysis is an exploratory tool and so it should be used to guide the researcher to make various decisions: you shouldn't leave the computer to make them. One important decision is the number of factors to extract (Section 17.4.5). By Kaiser's criterion we should extract four factors (which is what SPSS has done); however, this criterion is accurate when there are fewer than 30 variables and communalities after extraction are greater than .7, or when the sample size exceeds 250 and the average communality is greater than .6. No communalities exceed .7 (Output 17.5), and the average communality can be found by adding them up and dividing by the number of communalities (9.31/23 = .405). So, both of these criteria suggest Kaiser's rule might be inappropriate for these data. We could use Jolliffe's criterion (retain factors with eigenvalues greater than .7), but there is little to recommend this criterion over Kaiser's and we'd end up with 10 factors (see Output 17.4). Finally, we could use the scree plot, which we asked SPSS to produce by using the option in Figure 17.11. This curve is difficult to interpret because there are points of inflexion at both 3 and 5 factors (Output 17.6). Therefore, we could probably justify retaining either two or four factors.

So how many factors should we extract? We need to consider that the recommendations for Kaiser's criterion are for much smaller samples than we have. Therefore, given our huge sample, and given that there is some consistency between Kaiser's criterion and the scree plot, it is reasonable to extract four factors; however, you might like to rerun the analysis specifying that SPSS extract only two factors (see Figure 17.11) and compare the results.

Output 17.7 shows an edited version of the reproduced correlation matrix that was requested using the option in Figure 17.10. The top half of this matrix (labelled *Reproduced Correlations*) contains the correlation coefficients between all of the questions based on the factor model. The diagonal of this matrix contains the communalities after extraction for each variable (you can check the values against Output 17.5).



OUTPUT 17.6

Reproduced Correlations

		Question_01	Question_02	Question_03	Question_04	Question_05	Question_19	Question_20	Question_21	Question_22	Question_23
Reproduced Correlation	Question_01	.371	-.112	-.338	.393	.328	-.191	.266	.398	-.072	-.011
	Question_02	-.112	.260	.295	-.129	-.119	.237	-.192	-.201	.227	.146
	Question_03	-.338	.295	.472	-.367	-.316	.328	-.336	-.431	.242	.133
	Question_04	.393	-.129	-.367	.419	.353	-.214	.282	.429	-.092	-.021
	Question_05	.328	-.119	-.316	.353	.299	-.190	.237	.364	-.091	-.025
	Question_06	.221	-.076	-.218	.269	.249	-.167	.078	.259	-.175	-.072
	Question_07	.349	-.154	-.363	.393	.344	-.243	.230	.408	-.173	-.066
	Question_08	.345	-.044	-.258	.345	.277	-.129	.172	.285	-.086	-.055
	Question_09	-.071	.290	.295	-.092	-.092	.255	-.174	-.124	.272	.178
	Question_10	.191	-.096	-.210	.218	.194	-.149	.116	.223	-.130	-.061
	Question_11	.362	-.131	-.345	.375	.311	-.210	.213	.339	-.178	-.110
	Question_12	.374	-.189	-.407	.412	.356	-.265	.291	.447	-.158	-.057
	Question_13	.329	-.143	-.341	.371	.325	-.231	.202	.375	-.182	-.078
	Question_14	.342	-.155	-.359	.381	.333	-.238	.237	.400	-.160	-.061
	Question_15	.289	-.160	-.327	.319	.277	-.223	.204	.331	-.180	-.091
	Question_16	.401	-.193	-.426	.430	.364	-.267	.315	.457	-.152	-.063
	Question_17	.379	-.089	-.321	.393	.324	-.181	.212	.351	-.123	-.066
	Question_18	.355	-.155	-.369	.402	.354	-.249	.230	.419	-.179	-.066
	Question_19	-.191	.237	.328	-.214	-.190	.245	-.218	-.271	.211	.124
	Question_20	.266	-.192	-.336	.282	.237	-.218	.266	.329	-.122	-.059
	Question_21	.398	-.201	-.431	.429	.364	-.271	.329	.468	-.142	-.051
	Question_22	-.072	.227	.242	-.092	-.091	.211	-.122	-.142	.247	.163
	Question_23	-.011	.146	.133	-.021	-.025	.124	-.059	-.051	.163	.116
Residual[b]	Question_01		.013	.001	.042	.074	.002	-.052	-.003	-.032	.009
	Question_02	.013		.023	.017	-.001	-.034	-.010	-.004	.004	-.046
	Question_03	.001	.023		-.014	.006	.014	.011	.014	-.039	.017
	Question_04	.042	.017	-.014		.048	.028	-.039	-.016	-.006	-.013
	Question_05	.074	-.001	.006	.048		.025	-.037	-.030	-.041	-.017
	Question_06	-.004	.004	-.009	.009	.009	.000	.022	.013	.010	.003
	Question_07	-.044	-.006	-.019	.016	-.005	-.026	-.009	.075	.005	-.004
	Question_08	-.014	-.005	.000	.004	-.009	-.030	.003	.013	.006	.005
	Question_09	-.022	.024	.005	-.033	-.003	-.005	.015	.038	-.015	-.007
	Question_10	.023	.012	.017	-.003	.064	.022	-.032	-.030	-.001	-.001
	Question_11	-.005	-.013	-.006	-.007	-.013	.011	.042	.007	.016	.023
	Question_12	-.028	-.006	-.003	.030	-.009	-.001	.007	-.007	-.009	.011
	Question_13	.025	.000	.023	-.028	-.024	.004	.002	-.001	-.014	.025
	Question_14	-.004	-.009	-.012	-.030	-.017	-.016	-.011	-.001	-.009	.012
	Question_15	-.044	-.005	.015	.015	-.016	.013	.002	-.031	.012	.029
	Question_16	.098	.025	.007	-.014	.010	.000	-.050	-.036	-.003	-.019
	Question_17	-.009	.002	-.006	-.010	-.014	.018	-.007	.012	-.003	-.026
	Question_18	-.008	-.009	-.006	-.020	-.032	-.007	.005	.011	.019	-.014
	Question_19	.002	-.034	.014	.028	.025		-.031	-.004	.023	-.002
	Question_20	-.052	-.010	.011	-.039	-.037	-.031		.135	.022	.024
	Question_21	-.003	-.004	.014	-.018	-.030	-.004	.139		.013	-.017
	Question_22	-.032	.004	-.039	-.006	-.041	.023	.022	.015		.067
	Question_23	.009	-.046	.017	-.013	-.017	-.002	.024	-.017	.067	

Extraction Method: Principal Axis Factoring.

b. Residuals are computed between observed and reproduced correlations. There are 12 (4.0%) nonredundant residuals with absolute values greater than 0.05.

OUTPUT 17.7

The correlations in the reproduced matrix differ from those in the *R*-matrix because they stem from the model rather than the observed data. If the model were a perfect fit of the data then we would expect the reproduced correlation coefficients to be the same as the original correlation coefficients. Therefore, to assess the fit of the model we can look at the differences between the observed correlations and the correlations based on the model. For example, if we take the correlation between questions 1 and 2, the correlation based on the observed data is −.099 (taken from Output 17.1). The correlation based on the model is −.112, which is slightly higher. We can calculate the difference as follows:

$$\text{residual} = r_{\text{observed}} - r_{\text{from model}}$$

$$\text{residual}_{Q_1Q_2} = (-0.099) - (-0.112)$$

$$= 0.013$$

You should notice that this difference is the value quoted in the lower half of the reproduced matrix (labelled *Residual*) for questions 1 and 2 (highlighted in blue). Therefore, the lower half of the reproduced matrix contains the differences between the observed correlation coefficients and the ones predicted from the model. For a good model these values will all be small. In fact, we want most values to be less than .05. Rather than scan this huge matrix, SPSS provides a footnote summary, which states how many residuals have an absolute value greater than .05. For these data there are only 12 residuals (4%)[8] that are greater than .05. There are no hard-and-fast rules about what proportion of residuals should be below .05; however, if more than 50% are greater than .05 you probably have grounds for concern. For these data we have around 4%, which is certainly nothing to worry about.

CRAMMING SAM'S TIPS Factor extraction

- To decide how many factors to extract, look at the table labelled *Communalities* and the column labelled *Extraction*. If these values are all .7 or above and you have less than 30 variables then the SPSS default (Kaiser's criterion) for extracting factors is fine. Likewise, if your sample size exceeds 250 and the average of the communalities is .6 or greater then the default option is fine. Alternatively, with 200 or more participants the scree plot can be used.
- Check the bottom of the table labelled *Reproduced Correlations* for the percentage of 'nonredundant residuals with absolute values greater than 0.05'. This percentage should be less than 50% and the smaller it is, the better.

17.7.3. Factor rotation ②

The first analysis I asked you to run was using an orthogonal rotation. However, I also asked you to rerun the analysis using oblique rotation. In this section the results of both analyses will be reported so as to highlight the differences between the outputs. This comparison will also be a useful way to show the circumstances in which one type of rotation might be preferable to another.

17.7.3.1. Orthogonal rotation (varimax) ②

Output 17.8 shows the rotated factor matrix (called the rotated component matrix in PCA), which is a matrix of the factor loadings for each variable on each factor. This matrix contains the same information as the factor matrix in Output 17.5, except that it is calculated *after* rotation. There are several things to consider about the format of this matrix. First, factor loadings less than .3 have not been displayed because we asked for these loadings to be suppressed using the option in Figure 17.14. Second, the variables are listed in the order of size of their factor loadings because we asked for the output to be *Sorted by size* using the option in Figure 17.14. If this option was not selected the variables would be listed in the order they appear in the data editor. Finally, for all other parts of the output I suppressed the variable labels (to save space), but for this output I have used the variable labels to aid interpretation.

Compare this matrix to the unrotated solution (Output 17.5). Before rotation, most variables loaded highly on the first factor and the remaining factors didn't really get a look-in. However, the rotation of the factor structure has clarified things considerably: there are four factors and most variables load very highly on only one factor.[9] In cases where a variable loads highly on more than one factor the loading is typically higher for one factor than another. For example, 'SPSS always crashes when I try to use it' loads on both factor 1 and 2, but the loading for factor 2 (.612) is higher than for factor 1 (.366), so it makes sense to think of it as part of factor 2 more than factor 1. Remember that every variable has a loading on every factor, it just appears as though they don't in Output 17.8 because we asked that they not be printed if they were lower than .3.

The next step is to look at the content of questions that load highly on the same factor to try to identify common themes. If the mathematical factors represent some real-world construct then common themes among highly loading questions can help us identify what the construct might be. The questions that load highly on factor 1 seem to relate to different aspects of statistics; therefore, we might label this factor *fear of statistics*. The questions that load highly on factor 2 all seem to relate to using computers or SPSS. Therefore we might label this factor *fear of computers*. The three questions that load highly on factor 3 all seem to relate to mathematics; therefore, we might label this factor *fear of mathematics*. Finally, the questions that load highly on factor 4 contain some component of social evaluation from friends; therefore, we might label this factor *peer evaluation*. This analysis seems to reveal that the questionnaire is composed of four subscales: fear of statistics, fear of computers, fear of maths and fear of negative peer evaluation. There are two possibilities here. The first is that the SAQ failed to measure what it set out to (namely, SPSS anxiety) but does measure some related constructs. The second is that these four constructs are sub-components of SPSS anxiety; however, the factor analysis does not indicate which of these possibilities is true.

Rotated Factor Matrix[a]

	Factor			
	1	2	3	4
I wake up under my duvet thinking that I am trapped under a normal distribution	.594			
I weep openly at the mention of central tendency	.543			
I dream that Pearson is attacking me with correlation coefficients	.527			
People try to tell you that SPSS makes statistics easier to understand but it doesn't	.510	.398		
Standard deviations excite me	-.505			.399
Statistics makes me cry	.504			
I can't sleep for thoughts of eigenvectors	.465			
I don't understand statistics	.436			
I have little experience of computers		.753		
SPSS always crashes when I try to use it	.366	.612		
I worry that I will cause irreparable damage because of my incompetence with computers		.564		
All computers hate me	.364	.559		
Computers have minds of their own and deliberately go wrong whenever I use them	.388	.485		
Computers are useful only for playing games		.380		
Computers are out to get me		.377		
I have never been good at mathematics			.759	
I did badly at mathematics at school			.688	
I slip into a coma whenever I see an equation			.641	
My friends are better at statistics than me				.559
My friends are better at SPSS than I am				.465
My friends will think I'm stupid for not being able to cope with SPSS				.464
Everybody looks at me when I use SPSS				.375
If I'm good at statistics my friends will think I'm a nerd				.329

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 7 iterations.

OUTPUT 17.8

17.7.3.2. Oblique rotation ②

When an oblique rotation is conducted the factor matrix is split into two matrices: the *pattern matrix* and the *structure matrix* (see Jane Superbrain Box 17.1). For orthogonal rotation these matrices are the same. The pattern matrix contains the factor loadings and is comparable to the factor matrix that we interpreted for the orthogonal rotation. The structure matrix takes into account the relationship between factors (in fact it is a product of the pattern matrix and the matrix containing the correlation coefficients between factors). Most researchers interpret the pattern matrix, because it is usually simpler; however, there are situations in which values in the pattern matrix are suppressed because of relationships between the factors. Therefore, the structure matrix is a useful double-check and Graham et al. (2003) recommend reporting both (with some useful examples of why this can be important).

For the pattern matrix for these data (Output 17.9) the same four factors seem to have emerged. Factor 1 seems to represent fear of statistics, factor 2 represents fear of peer evaluation, factor 3 represents fear of computers and factor 4 represents fear of mathematics. The structure matrix (Output 17.10) differs in that shared variance is not ignored. The picture becomes more complicated because, with the exception of factor 2, several variables load highly on more than one factor. This has occurred because of the relationship between factors 1 and 3 and between factors 3 and 4. This example should highlight why the pattern matrix is preferable for interpretative reasons: it contains information about the *unique* contribution of a variable to a factor.

The final part of the output is a correlation matrix between the factors (Output 17.11). This matrix contains the correlation coefficients between factors. As predicted from the structure matrix, factor 2 has fairly small relationships with the other factors, but all other factors have fairly large correlations. The fact that these correlations exist tells us that the constructs measured can be interrelated. If the constructs were independent then we would expect oblique rotation to provide an identical solution to an orthogonal rotation and the factor correlation matrix should be an identity matrix (i.e., all factors have correlation coefficients of 0). Therefore, this matrix can be used to assess whether it is reasonable to assume independence between factors: for these data it appears that we cannot assume independence and so the obliquely rotated solution is probably a better representation of reality.

Pattern Matrix[a]

	Factor			
	1	2	3	4
I wake up under my duvet thinking that I am trapped under a normal distribution	.536			
I can't sleep for thoughts of eigenvectors	.470			
I weep openly at the mention of central tendency	.449			
I dream that Pearson is attacking me with correlation coefficients	.441			
Standard deviations excite me	-.435	.324		
Statistics makes me cry	.432			
People try to tell you that SPSS makes statistics easier to understand but it doesn't	.412		.358	
I don't understand statistics	.357			
My friends are better at statistics than me		.559		
My friends are better at SPSS than I am		.465		
My friends will think I'm stupid for not being able to cope with SPSS		.453		
If I'm good at statistics my friends will think I'm a nerd		.345		
Everybody looks at me when I use SPSS		.336		
I have little experience of computers			.862	
SPSS always crashes when I try to use it			.635	
All computers hate me			.562	
I worry that I will cause irreparable damage because of my incompetence with computers			.558	
Computers have minds of their own and deliberately go wrong whenever I use them			.473	
Computers are useful only for playing games			.386	
Computers are out to get me			.318	
I have never been good at mathematics				-.851
I did badly at mathematics at school				-.734
I slip into a coma whenever I see an equation				-.675

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.[a]

a. Rotation converged in 17 iterations.

OUTPUT 17.9

CRAMMING SAM'S TIPS Interpretation

- If you've conduced orthogonal rotation then look at the table labelled *Rotated Component Matrix*. For each variable, note the factor/component for which the variable has the highest loading (by 'high' I mean loadings above .4 when you ignore the plus or minus sign). Try to make sense of what the factors represent by looking for common themes in the items that load on them.
- If you've conducted oblique rotation then do the same as above but for the table labelled *Pattern Matrix*. Double-check what you find by doing the same thing for the structure matrix.

Structure Matrix

	Factor			
	1	2	3	4
I wake up under my duvet thinking that I am trapped under a normal distribution	.657		.475	-.391
I weep openly at the mention of central tendency	.621		.493	-.469
Standard deviations excite me	-.596	.486	-.409	.369
People try to tell you that SPSS makes statistics easier to understand but it doesn't	.593		.564	-.366
I dream that Pearson is attacking me with correlation coefficients	.586		.472	-.458
Statistics makes me cry	.552		.407	-.449
I can't sleep for thoughts of eigenvectors	.496			
I don't understand statistics	.492		.422	-.374
My friends are better at statistics than me		.572		
My friends will think I'm stupid for not being able to cope with SPSS		.486		
My friends are better at SPSS than I am		.484		
Everybody looks at me when I use SPSS	-.360	.425		
If I'm good at statistics my friends will think I'm a nerd		.328		
I have little experience of computers			.746	-.341
SPSS always crashes when I try to use it	.486		.720	-.407
All computers hate me	.479		.676	-.415
I worry that I will cause irreparable damage because of my incompetence with computers	.414		.673	-.457
Computers have minds of their own and deliberately go wrong whenever I use them	.489		.613	-.390
Computers are out to get me	.384		.510	-.428
Computers are useful only for playing games			.437	
I have never been good at mathematics	.314		.353	-.798
I did badly at mathematics at school	.369		.478	-.783
I slip into a coma whenever I see an equation	.404		.476	-.750

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.

OUTPUT 17.10

Factor Correlation Matrix

Factor	1	2	3	4
1	1.000	-.296	.483	-.429
2	-.296	1.000	-.302	.186
3	.483	-.302	1.000	-.532
4	-.429	.186	-.532	1.000

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.

OUTPUT 17.11

On a theoretical level the dependence between our factors does not cause concern; we might expect a fairly strong relationship between fear of maths, fear of statistics and fear of computers. Generally, the less mathematically and technically minded people struggle with statistics. However, we would not necessarily expect these constructs to correlate strongly with fear of peer evaluation (because this construct is more socially based). In fact, this factor is the one that correlates the least with all others – so, on a theoretical level, things have turned out rather well.

17.7.4. Factor scores ②

Having reached a suitable solution and rotated that solution, we can look at the factor scores. SPSS will display the component score matrix *B* (see Section 17.3.3.1) from which the factor scores are calculated. I haven't reproduced this table here because I can't think of a reason why most people would want to look at it. In the original analysis we asked for scores to be calculated based on the Anderson–Rubin method. You will find these scores in the data editor. There should be four new columns of data (one for each factor) labelled *FAC1_1*, *FAC2_1*, *FAC3_1* and *FAC4_1*, respectively. If you asked for factor scores in the oblique rotation then these scores will appear in the data editor in four other columns labelled *FAC2_1* and so on.

SELF-TEST Using what you learnt in Section 8.7.6, use the *Case Summaries command to list the factor scores for these data (given that there are over 2500 cases, you might like to restrict the output to the first 10).*

Case Summaries[a]

		A-R factor score 1 for analysis 1	A-R factor score 2 for analysis 1	A-R factor score 3 for analysis 1	A-R factor score 4 for analysis 1
1		-1.12974	.05090	-1.58646	-.55242
2		-.04484	-.47739	-.22126	.64055
3		.15620	-.72240	.08299	-.90901
4		.79370	.61178	-.79341	-.31779
5		-.98251	.66284	-.35819	.54788
6		-.59551	2.13562	-.53156	-.52313
7		-1.33140	-.19415	.08213	.87306
8		-.91760	-.20011	-.02149	.96984
9		1.70800	1.45700	3.03959	.65963
10		-.37637	-.77093	.06181	1.58454
Total	N	10	10	10	10

a. Limited to first 10 cases.

OUTPUT 17.12

Output 17.12 shows the factor scores for the first 10 participants. It should be pretty clear that participant 9 scored highly on factors 1 to 3 and so this person is very anxious about statistics, computing and maths, but less so about peer evaluation (factor 4). Factor scores can be used in this way to assess the relative fear of one person compared to another, or we could add the scores up to obtain a single score for each participant (which we might assume represents SPSS anxiety as a whole). We can also use factor scores in regression when groups of predictors correlate so highly that there is multicollinearity. However, people do not normally use factor scores themselves but instead sum scores on items that they have decided load on the same factor (e.g., create a score for statistics anxiety by adding up a person's scores on items 1, 3, 4, 5, 12, 16, 20 and 21).

17.7.5. Summary ②

To sum up, the analyses revealed four underlying scales in our questionnaire that may or may not relate to genuine sub-components of SPSS anxiety. It also seems as though an obliquely rotated solution was preferred due to the interrelationships between factors. The use of factor analysis is purely exploratory; it should be used only to guide future hypotheses, or to inform researchers about patterns within data sets. A great many decisions are left to the researcher using factor analysis and I urge you to make informed decisions, rather than basing decisions on the outcomes you would like to get. The next question is whether or not our scale is reliable.

17.8. How to report factor analysis ①

When reporting factor analysis we should provide our readers with enough information to form an informed opinion about what we've done. We should be clear about our criteria for extracting factors and the method of rotation used. We should also produce a table of the rotated factor loadings of all items and flag (in bold) values above a criterion level (I would personally choose .40, but see Section 17.4.6.2). We should also report the percentage of variance that each factor explains and possibly the eigenvalue too. Table 17.1 shows an example of such a table for the SAQ data (oblique rotation); note that I have also reported the sample size in the title.

In my opinion, a table of factor loadings and a description of the analysis are a bare minimum. You could consider (if it's not too large) including the table of correlations from which someone could reproduce your analysis (should they want to), and some information on sample size adequacy. For this example we might write something like this:

> A principal axis factor analysis was conducted on the 23 items with oblique rotation (direct oblimin). The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis, KMO = .93 ('marvellous' according to Hutcheson & Sofroniou, 1999), and all KMO values for individual items were greater than .77, which is well above the acceptable limit of .5 (Field, 2013). An initial analysis was run to obtain eigenvalues for each factor in the data. Four factors had eigenvalues over Kaiser's criterion of 1 and in combination explained 50.32% of the variance. The scree plot was ambiguous and showed inflexions that would justify retaining either 2 or 4 factors. We retained 4 factors because of the large sample size and the convergence of the scree plot and Kaiser's criterion on this value. Table 17.1 shows the factor loadings after rotation. The items that cluster on the same factor suggest that factor 1 represents a fear of statistics, factor 2 represents peer evaluation concerns, factor 3 a fear of computers and factor 4 a fear of maths.

17.9. Reliability analysis ②

17.9.1. Measures of reliability ③

If you're using factor analysis to validate a questionnaire, it is useful to check the reliability of your scale.

 SELF-TEST Thinking back to Chapter 1, what are reliability and test–retest reliability?

Reliability means that a measure (or in this case questionnaire) should consistently reflect the construct that it is measuring. One way to think of this is that, other things being equal, a person should get the same score on a questionnaire if they complete it at two different points in time (we have already discovered that this is called test–retest reliability). So, someone who is terrified of SPSS and who scores highly on our SAQ should score similarly highly if we tested them a month later (assuming they hadn't gone into some kind of SPSS-anxiety therapy in that month). Another way to look at reliability is to say that two people who are the same in terms of the construct being measured should get the same score. So, if we took two people who were equally SPSS-phobic, then they should get more or less identical scores on the SAQ. Likewise, if we took two people who loved SPSS, they should both get equally low scores. It should be apparent that the SAQ wouldn't be an accurate measure of SPSS anxiety if we took someone who loved SPSS and someone who was terrified of it and they got the same score! In statistical terms, the usual way to look at reliability is based on the idea that individual items (or sets of items) should produce results consistent with the overall questionnaire. So, if we take someone scared of SPSS, then their overall score on the SAQ will be high; if the SAQ is reliable then if we randomly select some items from it the person's score on those items should also be high.

TABLE 17.1 Summary of exploratory factor analysis results for the SPSS anxiety questionnaire (*N* = 2571)

Item	Rotated Factor Loadings: Fear of Statistics	Peer Evaluation	Fear of Computers	Fear of Maths
I wake up under my duvet thinking that I am trapped under a normal distribution	**.54**	−.04	.17	−.06
I can't sleep for thoughts of eigenvectors	**.47**	−.14	−.08	−.05
I weep openly at the mention of central tendency	**.45**	−.05	.17	−.18
I dream that Pearson is attacking me with correlation coefficients	**.44**	.08	.18	−.19
Standard deviations excite me	**−.44**	.32	−.05	.10
Statistics makes me cry	**.43**	.10	.11	−.23
People try to tell you that SPSS makes statistics easier to understand but it doesn't	**.41**	−.04	.36	.01
I don't understand statistics	**.36**	.05	.20	−.13
My friends are better at statistics than me	−.09	**.56**	−.02	−.11
My friends are better at SPSS than I am	.07	**.47**	−.11	.04
My friends will think I'm stupid for not being able to cope with SPSS	−.18	**.45**	.04	−.05
If I'm good at statistics my friends will think I'm a nerd	.10	**.35**	.00	.07
Everybody looks at me when I use SPSS	−.22	**.34**	−.08	.01
I have little experience of computers	−.22	−.01	**.86**	.03
SPSS always crashes when I try to use it	.18	−.01	**.64**	.01
All computers hate me	.19	−.02	**.56**	−.03
I worry that I will cause irreparable damage because of my incompetence with computers	.08	−.04	**.56**	−.12
Computers have minds of their own and deliberately go wrong whenever I use them	.24	−.02	**.47**	−.03
Computers are useful only for playing games	.00	−.06	**.39**	−.06
Computers are out to get me	.11	−.13	**.32**	−.19
I have never been good at mathematics	.01	.05	−.09	**−.85**
I did badly at mathematics at school	−.01	−.11	.06	**−.73**
I slip into a coma whenever I see an equation	.08	.02	.09	**−.68**
Eigenvalues	7.29	1.74	1.32	1.23
% of variance	31.70	7.56	5.73	5.34
α	.82	.57	.82	.82

Note: Factor loadings over .40 appear in bold

LABCOAT LENI'S REAL RESEARCH 17.1

Worldwide addiction? ②

In 2007 it was estimated that around 179 million people worldwide used the Internet. From the increasing popularity (and usefulness) of the Internet has emerged a serious and recognized problem of internet addiction. To research this construct it's helpful to be able to measure it, so Laura Nichols and Richard Nicki developed the Internet Addiction Scale (Nichols & Nicki, 2004). Nichols and Nicki's 36-item questionnaire contains items such as 'I have stayed on the Internet longer than I intended to' and 'My grades/work have suffered because of my Internet use' to which responses are made on a 5-point scale (Never, Rarely, Sometimes, Frequently, Always). (Incidentally, while researching this topic I encountered an Internet addiction recovery website that offered a whole host of resources (e.g., questionnaires, online support groups, videos, podcasts, etc.) that would keep you online for ages. It struck me that this was like having a heroin addiction recovery centre that had a huge pile of free heroin in the reception area.)

The data from 207 people in this study are in the file **Nichols & Nicki (2004).sav**. The authors dropped two items because they had low means and variances, and dropped three others because of relatively low correlations with other items. They performed a principal component analysis on the remaining 31 items. Labcoat Leni wants you to run some descriptive statistics to work out which two items were dropped for having low means/variances, then inspect a correlation matrix to find the three items that were dropped for having low correlations. Finally, he wants you to run a principal component analysis on the data. Answers are in the additional material on the companion website (or look at the original article).

NICHOLS, L. A., & NICKI, R. (2004). *PSYCHOLOGY OF ADDICTIVE BEHAVIORS* , 18 Ã , 381–384.

The simplest way to do this in practice is to use **split-half reliability**. This method splits the scale set into two randomly selected sets of items. A score for each participant is calculated on each half of the scale. If a scale is reliable a person's score on one half of the scale should be the same (or similar) to their score on the other half. Across several participants, scores from the two halves of the questionnaire should correlate very highly. The correlation between the two halves is the statistic computed in the split-half method, with large correlations being a sign of reliability. The problem with this method is that there are several ways in which a set of data can be randomly split into two and so the results could be a product of the way in which the data were split. To overcome this problem, Cronbach (1951) came up with a measure that is loosely equivalent to creating two sets of items in every way possible and computing the correlation coefficient for each split. The average of these values is equivalent to **Cronbach's alpha,** α, which is the most common measure of scale reliability:[10]

$$\alpha = \frac{N^2\overline{\text{cov}}}{\sum s^2_{\text{item}} + \sum \text{cov}_{\text{item}}} \tag{17.6}$$

This equation may look complicated, but actually isn't. For each item on our scale we can calculate two things: the variance within the item, and the covariance between a particular item and any other item on the scale. Put another way, we can construct a variance–covariance matrix of all items. In this matrix the diagonal elements will be the variance within a particular item, and the off-diagonal elements will be covariances between pairs of items. The top half of the equation is simply the number of items (N) squared multiplied by the average covariance between items (the average of the off-diagonal elements in the aforementioned variance–covariance matrix). The bottom half is the sum of all the item variances and item covariances (i.e., the sum of everything in the variance–covariance matrix).

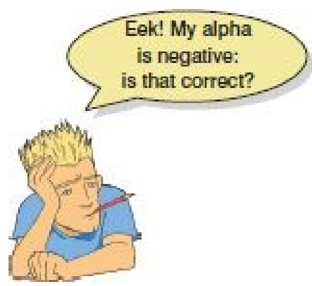
There is a standardized version of the coefficient too, which essentially uses the same equation except that correlations are used rather than covariances, and the bottom half of the equation uses the sum of the elements in the correlation matrix of items (including the 1s that appear on the diagonal of that matrix). The normal alpha is appropriate when items on a scale are summed to produce a single score for that scale (the standardized alpha is not appropriate in these cases). The standardized alpha is useful, though, when items on a scale are standardized before being summed.

17.9.2. Interpreting Cronbach's α (some cautionary tales) ②

You'll often see in books or journal articles, or be told by people, that a value of .7 to .8 is an acceptable value for Cronbach's α; values substantially lower indicate an unreliable scale. Kline (1999) notes that although the generally accepted value of .8 is appropriate for cognitive tests such as intelligence tests, for ability tests a cut-off point of .7 is more suitable. He goes on to say that when dealing with psychological constructs, values below even .7 can, realistically, be expected because of the diversity of the constructs being measured. Some even suggest that in the early stages of research, values as low as .5 will suffice (Nunnally, 1978). However, there are many reasons not to use these general guidelines, not least of which is that they distract you from thinking about what the value means within the context of the research you're doing (Pedhazur & Schmelkin, 1991).

We'll now look at some issues in interpreting alpha, which have been discussed particularly well by Cortina (1993) and Pedhazur and Schmelkin (1991). First, the value of α depends on the number of items on the scale. You'll notice that the top half of the equation for α includes the number of items squared. Therefore, as the number of items on the scale increases, α will increase. As such, it's possible to get a large value of α because you have a lot of items on the scale, and not because your scale is reliable. For example, Cortina (1993) reports data from two scales, both of which have α = .8. The first scale has only three items, and the average correlation between items was a respectable .57; however, the second scale had 10 items with an average correlation between these items of a less respectable .28. Clearly the internal consistency of these scales differs, but according to Cronbach's α they are both equally reliable.

Second, people tend to think that alpha measures 'unidimensionality', or the extent to which the scale measures one underlying factor or construct. This is true when there is one factor underlying the data (see Cortina, 1993), but Grayson (2004) demonstrates that data sets with the same α can have very different factor structures. He showed that α =.8 can be achieved in a scale with one underlying factor, with two moderately correlated factors and with two uncorrelated factors. Cortina (1993) has also shown that with more than 12 items, and fairly high correlations between items ($r > .5$), α can reach values around and above .7 (.65 to .84). These results show that α should not be used as a measure of 'uni-dimensionality'. Indeed, Cronbach (1951) suggested that if several factors exist then the formula should be applied separately to items relating to different factors. In other words, if your questionnaire has subscales, α should be applied separately to these subscales.



The final warning is about items that have a reverse phrasing. For example, in the SAQ there is one item (question 3) that was phrased the opposite way around to all other items. The item was 'standard deviations excite me'. Compare this to any other item and you'll see it requires the opposite response. For example, item 1 is 'statistics make me cry'. If you don't like statistics then you'll strongly agree with this statement and so will get a score of 5 on our scale. For item 3, if you hate statistics then standard deviations are unlikely to excite you so you'll strongly disagree and get a score of 1 on the scale. These reverse-phrased items are important for reducing response bias; participants will need to pay attention to the questions. For factor analysis, this reverse phrasing doesn't matter; all that happens is you get a negative factor loading for any reversed items (in fact, you'll see that item 3 has a negative factor loading in Output 17.9). However, these reverse-scored items will affect alpha. To see why, think about the equation for Cronbach's α. The top half incorporates the *average* covariance between items. If an item is reverse-phrased then it will have a negative relationship with other items, hence the covariances between this item and other items will be negative. The average covariance is the sum of covariances divided by the number of covariances, and by including a bunch of negative values we reduce the sum of covariances, and hence we also reduce Cronbach's α, because the top half of the equation gets smaller. In extreme cases, it is even possible to get a negative value for Cronbach's α, simply because the magnitude of negative covariances is bigger than the magnitude of positive ones. A negative Cronbach's α doesn't make much sense, but it does happen, and if it does, ask yourself whether you included any reverse-phrased items.

If you have reverse-phrased items then you also have to reverse the way in which they're scored before you conduct reliability analysis. This is quite easy. To take our SAQ data, we have one item which is currently scored as 1 = strongly disagree, 2 = disagree, 3 = neither, 4 = agree and 5 = strongly agree. This is fine for items phrased in such a way that agreement indicates statistics anxiety, but for item 3 (standard deviations excite me), disagreement indicates statistics anxiety. To reflect this numerically, we need to reverse the scale such that 1 = strongly agree, 2 = agree, 3 = neither, 4 = disagree and 5 = strongly disagree. In doing so, an anxious person still gets 5 on this item (because they'd strongly disagree with it).

To reverse the scoring find the maximum value of your response scale (in this case 5) and add 1 to it (so you get 6 in this case). Then for each person, you take this value and subtract from it the score they actually got. Therefore, someone who scored 5 originally now scores 6−5 = 1, and someone who scored 1 originally now gets 6−1 = 5. Someone in the middle of the scale with a score of 3 will still get 6−3 = 3. Obviously it would take a long time to do this for each person, but we can get SPSS to do it for us.

SELF-TEST Using what you learnt in Chapter 5, use the *compute* compute command to reverse-score item 3. (Clue: Remember that you are simply changing the variable to 6 minus its original value.)

17.9.3. Reliability analysis in SPSS ②

Let's test the reliability of the SAQ using the data in **SAQ.sav**. You should have reverse-scored item 3 (see above), but if you can't be bothered then load the file **SAQ (Item 3 Reversed).sav** instead. Remember also that I said we should conduct reliability analysis on any subscales individually. If we use the results from our oblique rotation (Output 17.9), then we have four subscales:

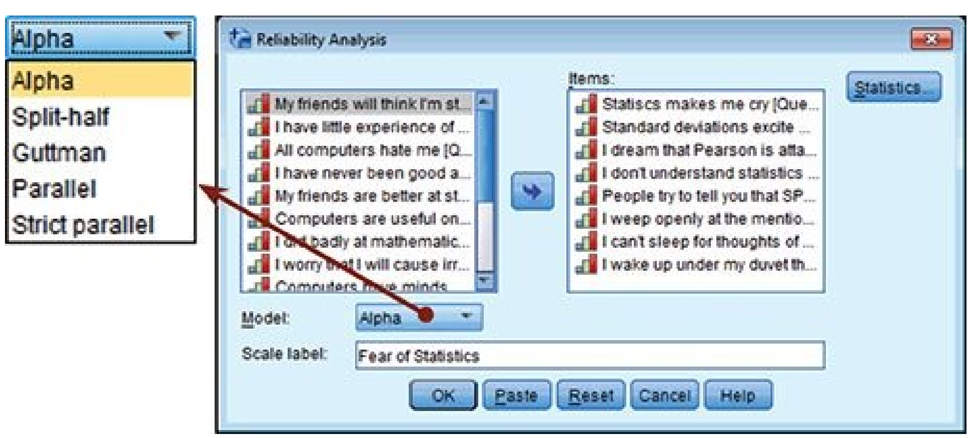


FIGURE 17.15 Main dialog box for reliability analysis.

① Subscale 1 (*Fear of statistics*): items 1, 3, 4, 5, 12, 16, 20, 21
② Subscale 2 (*Peer evaluation*): items 2, 9, 19, 22, 23
③ Subscale 3 (*Fear of computers*): items 6, 7, 10, 13, 14, 15, 18
④ Subscale 4 (*Fear of mathematics*): items 8, 11, 17

To conduct each reliability analysis on these data you need to select Analyze > Scale > Reliability Analysis... to display the dialog box in Figure 17.15. Select any items from the list that you want to analyse (to begin with, let's do the items from the fear of statistics subscale: items 1, 3, 4, 5, 12, 16, 20 and 21) on the left-hand side of the dialog box and drag them to the box labelled Items (or click on ➡). Remember that you can select several items at the same time if you hold down the *Ctrl* (*Cmd* on a Mac) key while you select the variables.

There are several reliability analyses you can run, but the default option is Cronbach's α. You can change the method (e.g., to the split-half method) by clicking on Alpha ▾ to reveal a drop-down list of possibilities, but the default method is a good one to select. Also, it's a good idea to type the name of the scale (in this case 'Fear of Statistics') into the box labelled *Scale label* because this will add a header to the SPSS output with whatever you type in this box: typing a sensible name here will make your output easier to follow.

If you click on Statistics... you can access the dialog box in Figure 17.16. In the statistics dialog box you can select several things, but the one most important for questionnaire reliability is: *Scale if item deleted*. This option tells us what the value of α would be if each item were deleted. If our questionnaire is reliable then we would not expect any one item to greatly affect the overall reliability. In other words, no item should cause a substantial decrease in α. If it does then you should consider dropping that item from the questionnaire to improve reliability.

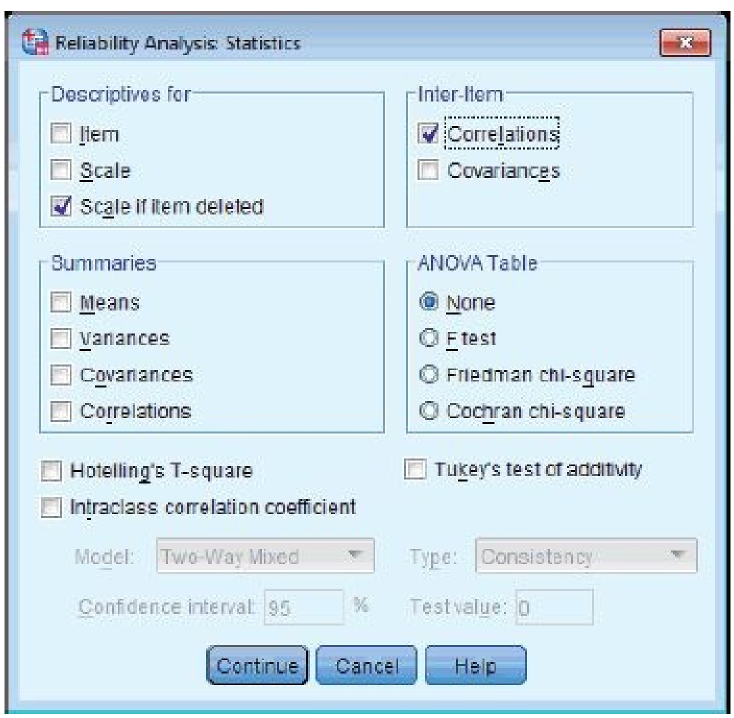


FIGURE 17.16
Statistics for reliability analysis

The inter-item correlations and covariances (and summaries) provide us with correlation coefficients and averages for items on our scale. We should already have these values from our factor analysis, so there is little point in selecting these options. Options like the *F test*, *Friedman chi-square* (if your data are ranked), *Cochran chi-square* (if your data are dichotomous), and *Hotelling's T-square* use these tests to compare the central tendency of different items on the questionnaire. These tests might be useful to check that items have similar distributional properties (i.e., the same average value), but given the large sample sizes you ought to be using for factor analysis, they will inevitably produce significant results even when only small differences exist between the questionnaire items.

You can also request an **intraclass correlation coefficient (ICC)**. The correlation coefficients that we encountered earlier in this book measure the relation between variables that measure different things. For example, the correlation between listening to Deathspell Omega and Satanism involves two classes of measures: the type of music a person likes and their religious beliefs. Intraclass correlations measure the relationship between two variables that measure the same thing (i.e., variables within the same class). Two common uses are in comparing paired data (such as twins) on the same measure, and assessing the consistency between judges' ratings of a set of objects (hence the reason why it is found in the reliability statistics in SPSS). If you'd like to know more, see Section 20.2.1.

Use the simple set of options in Figure 17.16 to run a basic reliability analysis. Click on Continue to return to the main dialog box and then click on OK to run the analysis.

17.9.4. Reliability analysis output ②

Output 17.13 shows the results of this basic reliability analysis for the fear of statistics subscale. The value of Cronbach's α is presented in a small table and indicates the overall reliability of the scale. Bearing in mind what we've already noted about effects from the number of items, and how daft it is to apply general rules, we're looking for values in the region of about .7 to .8. In this case α is .821, which is certainly in the region indicated by Kline (1999), and probably indicates good reliability.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Statistics makes me cry	21.76	21.442	.536	.343	.802
Standard deviations excite me	20.72	19.825	.549	.309	.800
I dream that Pearson is attacking me with correlation coefficients	21.35	20.410	.575	.355	.796
I don't understand statistics	21.41	20.942	.494	.272	.807
People try to tell you that SPSS makes statistics easier to understand but it doesn't	20.97	20.639	.572	.337	.796
I weep openly at the mention of central tendency	21.25	20.451	.597	.389	.793
I can't sleep for thoughts of eigenvectors	20.51	21.176	.419	.244	.818
I wake up under my duvet thinking that I am trapped under a normal distribution	20.96	19.939	.606	.399	.791

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.821	.823	8

OUTPUT 17.13

In the table labelled *Item-Total Statistics* the column labelled *Corrected Item-Total Correlation* has the correlations between each item and the total score from the questionnaire. In a reliable scale all items should correlate with the total. So, we're looking for items that don't correlate with the overall score from the scale: if any of these values are less than about .3 then we've got problems, because it means that a particular item does not correlate very well with the scale overall. Items with low correlations may have to be dropped. For these data, all data have item–total correlations above .3, which is encouraging.

The values in the column labelled *Cronbach's Alpha if Item Deleted* are the values of the overall α if that item isn't included in the calculation. As such, they reflect the change in Cronbach's α that would be seen if a particular item were deleted. The overall α is .821, and so all values in this column should be around that same value. We're actually looking for values of alpha greater than the overall α. If you think about it, if the deletion of an item increases Cronbach's α then this means that the deletion of that item improves reliability. Therefore, any items that have values of α in this column greater than the overall α may need to be deleted from the scale to improve its reliability. None of the items here would increase alpha if they were deleted, which is good news. It's worth noting that if items do need to be removed at this stage then you should rerun your factor analysis as well to make sure that the deletion of the item has not affected the factor structure

---

SELF-TEST Run reliability analyses on the other three subscales.

Just to illustrate the importance of reverse-scoring items before running reliability analysis, Output 17.14 shows the reliability analysis for the fear of statistics subscale but done on the original data (i.e., without item 3 being reverse-scored). Note that the overall α is considerably lower (.605 rather than .821). Also, note that this item has a negative item–total correlation (which is a good way to spot if you have a potential reverse-scored item in the data that hasn't been reverse-scored). Finally, note that for item 3, the α if item deleted is .8. That is, if this item were deleted then the reliability would improve from about .6 to about .8. This, I hope, illustrates that failing to reverse-score items that have been phrased oppositely to other items on the scale will mess up your reliability analysis.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Statistics makes me cry	20.93	12.125	.505	.343	.521
Standard deviations excite me	20.72	19.825	-.549	.309	.800
I dream that Pearson is attacking me with correlation coefficients	20.52	11.447	.526	.355	.505
I don't understand statistics	20.58	11.714	.466	.272	.523
People try to tell you that SPSS makes statistics easier to understand but it doesn't	20.14	11.739	.501	.337	.515
I weep openly at the mention of central tendency	20.42	11.584	.529	.389	.507
I can't sleep for thoughts of eigenvectors	19.68	12.107	.353	.244	.558
I wake up under my duvet thinking that I am trapped under a normal distribution	20.13	11.189	.541	.399	.497

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.605	.641	8

OUTPUT 17.14

Let's now look at our subscale of peer evaluation. For our subscale of peer evaluation you should get the output in Output 17.15. The overall reliability is .57, which is nothing to bake a cake for. The overall α is quite low, and although this is in keeping with what Kline says we should expect for this kind of social science data, it is well below the statistics subscale and (as we shall see) the other two. The scale has five items, compared to seven, eight and three on the other scales, so its reliability relative to the other scales is not going to be dramatically affected by the number of items. The values in the column labelled *Corrected Item-Total Correlation* are all around .3, and smaller for item 23. These results again indicate questionable internal consistency and identify item 23 as a potential problem. The values in the column labelled *Cronbach's Alpha if Item Deleted* indicate that none of the items here would increase the reliability if they were deleted because all values in this column are less than the overall reliability of .57. The items on this subscale cover quite diverse themes of peer evaluation, and this might explain the relative lack of consistency; we probably need to rethink this subscale.

Moving on to the fear of computers subscale, Output 17.16 shows an overall α of .823, which is pretty good. The values in the column labelled *Corrected Item-Total Correlation* are again all above .3, which is also good. The values in the column labelled *Cronbach's Alpha if Item Deleted* show that none of the items would increase the reliability if they were deleted. This indicates that all items are positively contributing to the overall reliability.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
My friends will think I'm stupid for not being able to cope with SPSS	11.46	8.119	.339	.134	.515
My friends are better at statistics than me	10.24	6.395	.391	.167	.476
Everybody looks at me when I use SPSS	10.79	7.381	.316	.106	.522
My friends are better at SPSS than I am	10.20	7.282	.378	.144	.487
If I'm good at statistics my friends will think I'm a nerd	9.65	7.988	.239	.069	.563

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.570	.572	5

OUTPUT 17.15

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
I have little experience of computers	15.87	17.614	.619	.398	.791
All computers hate me	15.17	17.737	.619	.395	.790
Computers are useful only for playing games	15.81	20.736	.400	.167	.824
I worry that I will cause irreparable damage because of my incompetenece with computers	15.64	18.809	.607	.384	.794
Computers have minds of their own and deliberately go wrong whenever I use them	15.22	18.719	.577	.350	.798
Computers are out to get me	15.33	19.322	.491	.250	.812
SPSS always crashes when I try to use it	15.52	17.832	.647	.447	.786

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.823	.821	7

OUTPUT 17.16

Finally, for the fear of maths subscale, Output 17.17 shows an overall reliability of .819, which indicates good reliability. The values in the column labelled *Corrected Item-Total Correlation* are all above .3, which is good, and the values in the column labelled *Cronbach's Alpha if Item Deleted* indicate that none of the items here would increase the reliability if they were deleted because all values in this column are less than the overall reliability value.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
I have never been good at mathematics	4.72	2.470	.684	.470	.740
I did badly at mathematics at school	4.70	2.453	.682	.467	.742
I slip into a coma whenever I see an equation	4.49	2.504	.652	.425	.772

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.819	.819	3

OUTPUT 17.17

CRAMMING SAM'S TIPS **Reliability**

- Reliability analysis is used to measure the consistency of a measure.
- Remember to reverse-score any items that were reverse-phrased on the original questionnaire before you run the analysis.
- Run separate reliability analyses for all subscales of your questionnaire.
- Cronbach's α indicates the overall reliability of a questionnaire, and values around .8 are good (or .7 for ability tests and the like).
- The *Cronbach's Alpha if Item Deleted* column tells you whether removing an item will improve the overall reliability. Values greater than the overall reliability indicate that removing that item will improve the overall reliability of the scale. Look for items that dramatically increase the value of α and remove them.
- If you remove items, rerun your factor analysis to check that the factor structure still holds.

17.10. How to report reliability analysis ②

You can report the reliabilities in the text using the symbol α and remembering that because Cronbach's α can't be larger than 1 we drop the zero before the decimal place (if we are following APA practice):

✓ The fear of computers, fear of statistics and fear of maths subscales of the SAQ all had high reliabilities, all Cronbach's α = .82. However, the fear of negative peer evaluation subscale had relatively low reliability, Cronbach's α = .57.

However, the most common way to report reliability analysis when it follows a factor analysis is to report the values of Cronbach's α as part of the table of factor loadings. For example, in Table 17.1 notice that in the last row of the table I quoted the value of Cronbach's α for each subscale in turn.

17.11. Brian's attempt to woo Jane ①

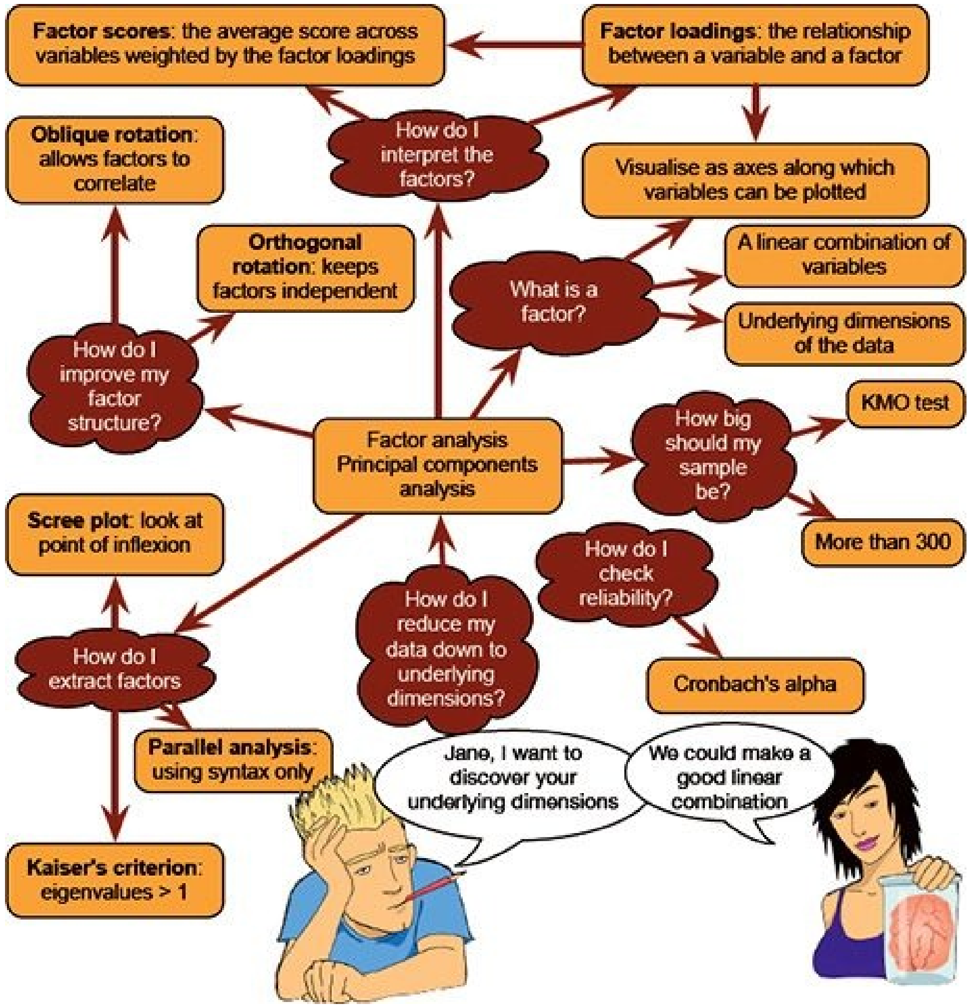


FIGURE 17.17
What Brian learnt from this chapter

17.12. What next? ②

At the age of 23 I took it upon myself to become a living homage to the digestive system. I furiously devoured articles and books on statistics (some of them I even understood), I mentally chewed over them, I broke them down with the stomach acid of my intellect, I stripped them of their goodness and nutrients, I compacted them down, and after about two years I forced the smelly brown remnants of those intellectual meals out of me in the form of a book. I was mentally exhausted at the end of it. ‘It’s a good job I’ll never have to do that again’, I thought.

17.13. Key terms that I’ve discovered

Alpha factoring
Anderson–Rubin method
Common factor
Common variance
Communality
Component matrix
Confirmatory factor analysis

Cronbach's α
Direct oblimin
Extraction
Equamax
Factor analysis
Factor loading
Factor matrix
Factor scores
Factor transformation matrix, Λ
Intraclass correlation coefficient (ICC)
Kaiser's criterion
Latent variable
Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy
Oblique rotation
Orthogonal rotation
Pattern matrix
Principal component analysis (PCA)
Promax
Quartimax
Random variance
Rotation
Scree plot
Singularity
Split-half reliability
Structure matrix
Unique factor
Unique variance
Varimax

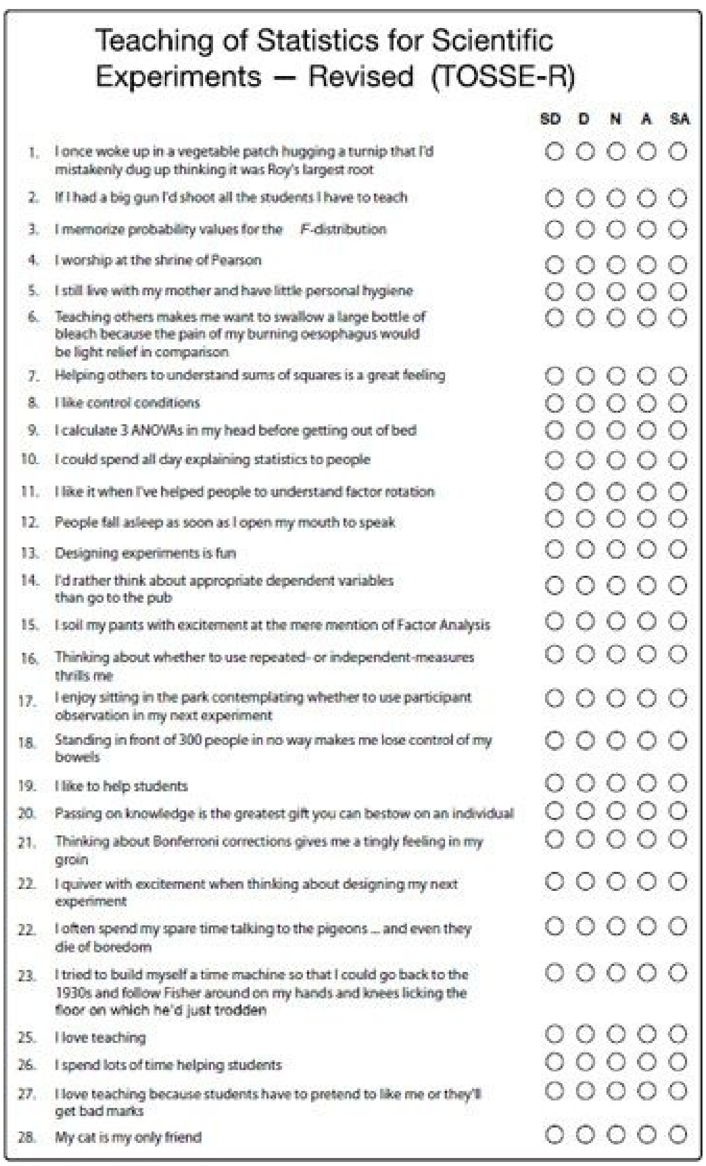
17.14. Smart Alex's tasks

- **Task 1**: Rerun the analysis in this chapter using principal component analysis and compare the results to those in the chapter. (Set the iterations to convergence to 30.) ②
- **Task 2**: The University of Sussex constantly seeks to employ the best people possible as lecturers. They wanted to revise the 'Teaching of Statistics for Scientific Experiments' (TOSSE) questionnaire, which is based on Bland's theory that says that good research methods lecturers should have: (1) a profound love of statistics; (2) an enthusiasm for experimental design; (3) a love of teaching; and (4) a complete absence of normal interpersonal skills. These characteristics should be related (i.e., correlated). The University revised this questionnaire to become the 'Teaching of Statistics for Scientific Experiments – Revised' (TOSSE-R). They gave this questionnaire to 239

research methods lecturers around the world to see if it supported Bland’s theory. The questionnaire is in Figure 17.18, and the data are in **TOSSE-R.sav**. Conduct a factor analysis (with appropriate rotation) and interpret the factor structure. ②

- **Task 3:** Dr Sian Williams (University of Brighton) devised a questionnaire to measure organizational ability. She predicted five factors to do with organizational ability: (1) preference for organization; (2) goal achievement; (3) planning approach; (4) acceptance of delays; and (5) preference for routine. These dimensions are theoretically independent. Williams’ questionnaire contains 28 items using a 7-point Likert scale (1 = strongly disagree, 4 = neither, 7 = strongly agree). She gave it to 239 people. Run a principal component analysis on the data in **Williams.sav**. ②
- **Task 4**: Zibarras, Port, and Woods (2008) looked at the relationship between personality and creativity. They used the Hogan Development Survey (HDS), which measures 11 dysfunctional dispositions of employed adults: being **volatile**, **mistrustful**, **cautious**, **detached**, **passive-aggressive**, **arrogant**, **manipulative**, **dramatic**, **eccentric**, **perfectionist**, and **dependent**. Zibarras et al. wanted to reduce these 11 traits and, based on parallel analysis, found that they could be reduced to three components. They ran a principal component analysis with varimax rotation. Repeat this analysis (**Zibarras et al. (2008).sav**) to see which personality dimensions clustered together (see page 210 of the original paper). ②

Answers can be found on the companion website.



Teaching of Statistics for Scientific Experiments — Revised (TOSSE-R)

	Item	SD	D	N	A	SA
1.	I once woke up in a vegetable patch hugging a turnip that I'd mistakenly dug up thinking it was Roy's largest root	○	○	○	○	○
2.	If I had a big gun I'd shoot all the students I have to teach	○	○	○	○	○
3.	I memorize probability values for the F-distribution	○	○	○	○	○
4.	I worship at the shrine of Pearson	○	○	○	○	○
5.	I still live with my mother and have little personal hygiene	○	○	○	○	○
6.	Teaching others makes me want to swallow a large bottle of bleach because the pain of my burning oesophagus would be light relief in comparison	○	○	○	○	○
7.	Helping others to understand sums of squares is a great feeling	○	○	○	○	○
8.	I like control conditions	○	○	○	○	○
9.	I calculate 3 ANOVAs in my head before getting out of bed	○	○	○	○	○
10.	I could spend all day explaining statistics to people	○	○	○	○	○
11.	I like it when I've helped people to understand factor rotation	○	○	○	○	○
12.	People fall asleep as soon as I open my mouth to speak	○	○	○	○	○
13.	Designing experiments is fun	○	○	○	○	○
14.	I'd rather think about appropriate dependent variables than go to the pub	○	○	○	○	○
15.	I soil my pants with excitement at the mere mention of Factor Analysis	○	○	○	○	○
16.	Thinking about whether to use repeated- or independent-measures thrills me	○	○	○	○	○
17.	I enjoy sitting in the park contemplating whether to use participant observation in my next experiment	○	○	○	○	○
18.	Standing in front of 300 people in no way makes me lose control of my bowels	○	○	○	○	○
19.	I like to help students	○	○	○	○	○
20.	Passing on knowledge is the greatest gift you can bestow on an individual	○	○	○	○	○
21.	Thinking about Bonferroni corrections gives me a tingly feeling in my groin	○	○	○	○	○
22.	I quiver with excitement when thinking about designing my next experiment	○	○	○	○	○
22.	I often spend my spare time talking to the pigeons ... and even they die of boredom	○	○	○	○	○
23.	I tried to build myself a time machine so that I could go back to the 1930s and follow Fisher around on my hands and knees licking the floor on which he'd just trodden	○	○	○	○	○
25.	I love teaching	○	○	○	○	○
26.	I spend lots of time helping students	○	○	○	○	○
27.	I love teaching because students have to pretend to like me or they'll get bad marks	○	○	○	○	○
28.	My cat is my only friend	○	○	○	○	○

FIGURE 17.18

The TOSSE-R questionnaire

17.15. Further reading

ʼtina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104. (A very readable paper on Cronbach's α.)

ıteman, G. E. (1989). *Principal components analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-069. Newbury Park, CA: Sage. (This monograph is quite high level but comprehensive.)

ıhazur, E., & Schmelkin, L. (1991). *Measurement, design and analysis*. Hillsdale, NJ: Erlbaum. (Chapter 22 is an excellent introduction to the theory of factor analysis.)

ıachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston: Allyn & Bacon.

---

[1] She didn't say 'rabbit', but she did say a word that describes what rabbits do a lot; it begins with an 'f' and the publishers think that it will offend you.

[2] PCA is not the same as factor analysis. This doesn't stop idiots like me from discussing them as though they are. I tend to focus on the similarities between the techniques, which will reduce some statisticians (and psychologists) to tears. I'm banking on these people not needing to read this book, so I'll take my chances because I think it's easier for you if I give you a general sense of what the procedures do and not obsess too much about their differences. Once you have got the basics under your belt, feel free to obsess about their differences and complain to all of your friends about how awful the book by that imbecile Field is …

[3] This matrix is called an *R*-matrix, or *R*, because it contains correlation coefficients and *r* usually denotes Pearson's correlation (see Chapter 7) – the *r* turns into a capital letter when it denotes a matrix.

[4] In his original paper Cattell advised including the factor at the point of inflexion as well, because it represents an error factor, or 'garbage can' as he put it. However, Thurstone argued that it is better to retain too few than too many factors, and in practice the 'garbage can' factor is rarely retained.

[5] This term means that the axes are at right angles to one another.

[6] To save space only columns for the first five and last five questions in the questionnaire are included.

[7] Actually the determinant of this matrix is 0.0005271; I have no idea why SPSS reports this value as .001.

[8] SPSS has a weird rounding habit here. There are 253 unique correlation coefficients in the table and 12 residuals greater than .05, which is (12/253) × 100 = 4.74%. SPSS seems to round down to the nearest whole percentage value for some reason.

[9] The suppression of loadings less than .3 and ordering variables by their loading size makes this pattern really easy to see.

[10] Although this is the easiest way to conceptualize Cronbach's, α, whether or not it is exactly equal to the average of all possible split-half reliabilities depends on exactly how you calculate the split-half reliability (see the glossary for computational details). If you use the Spearman–Brown formula, which takes no account of item standard deviations, then Cronbach's will be equal to the average split-half reliability only when the item standard deviations are equal; otherwise α will be smaller than the average. However, if you use a formula for split-half reliability that does account for item standard deviations (such as Flanagan, 1937; Rulon, 1939) then α will always equal the average split-half reliability (see Cortina, 1993).