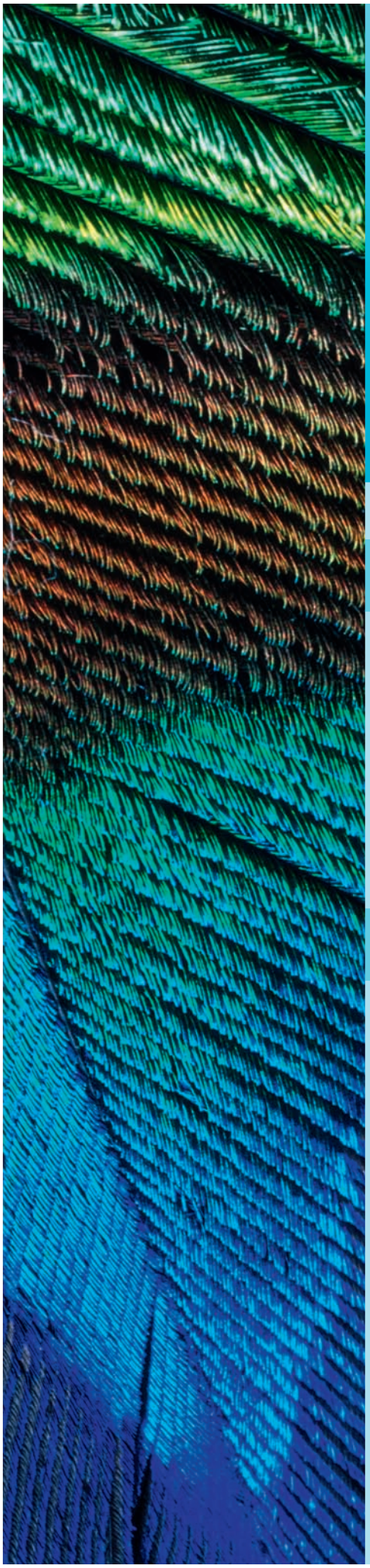


Chapter 25
An Introduction to Psychometric Testing

Key themes

- Psychometric testing
- Item writing
- Reliability
- Validity
- Exploratory factor analysis
- Confirmatory factor analysis
- Types and uses of psychometric tests

Learning outcomes

At the end of the chapter you should:

- Understand some of the ideas and criteria surrounding psychometric testing
- Know what is meant by reliability in psychometric testing
- Be able to explain what is meant by validity in psychometric testing
- How factor analysis is used in psychometric testing
- The types and uses for different psychometric tests

Introduction

When you leave university, you may be attending a number of job interviews. You may be aware that many employers use psychometric tests as part of the interview process for applicants. You may not be aware that in the spring of 2004, it was announced that personality tests designed to weed out racist applicants to the police were introduced in all 43 police forces in England and Wales. This was in response to a BBC documentary, *The Secret Policeman,* which exposed racism in the Greater Manchester police force. You may also not be aware that, as the *Sunday Telegraph* reported in October 2003, even James Murdoch, son of media tycoon Rupert Murdoch, was reported to have to sit psychometric tests to find out if he was fit to head his father's company, satellite broadcaster BSkyB. In March 2003, the UK higher education minister Margaret Hodge pointed to the diverse ways that universities could develop different forms of assessment by which to improve decisions on student's admission to university, and she encouraged universities simply to look beyond candidates' exams. Margaret Hodge argued that, like employers, universities could use a whole range of techniques including psychometric tests.

There is little doubt in our minds that measurement is a cornerstone of modern psychology. Even the government seems to be sanctioning measurement as an important aspect of education and work. Psychological tests are of fundamental importance to research in many areas of psychology. Probably the three areas in which that statement is most true are individual differences (including personality) and intelligence. However, psychological tests are also of immense value in developing areas such as health psychology (with its growing emphasis on quality of life), work psychology and educational psychology, as well as in more traditional areas such as social, cognitive and developmental psychology.

What we are going to do in this chapter, then, is introduce you to what makes a good psychometric test. There are some very simple but elegant ideas behind a good test. Namely, a test should be both reliable (that is, those items within measures correlate and sometimes are consistent over time) and valid (that is, the test measures what it claims to measure).

Source: Corbis/Tom Stewart

You will have already been introduced to the terms 'reliability' and 'validity' in research methods classes. Here, you will see how these ideas are central to psychometric testing.

Types and uses of psychometric tests

In your career as a psychologist, you will come across a variety of psychometric tests. The main types of tests you will come across in the personality, intelligence and individual differences literature are measures of personality, ability, motivation, educational and psychological work, clinical assessment and attitude.

- Personality measures are designed to measure a set of psychological traits or characteristics, of the person that remain relatively stable over time. An example of a personality measure might be a measure of the five-factor model of personality that would include questions and response choices that look for underlying tendencies of the person's behaviour.
- Ability measures are designed to measure particular abilities. These test often include intelligence tests (that measure an individual's ability at a number of cognitive processes such as perceptual speed, comprehension or reasoning) and Aptitude test (measurement of particular specific skills suited for a particularly task. Therefore, ability tests could include the measures of general intelligence,

or creative thinking, or successful communication strategies in the workplace. Items from ability tests seek to test these thought processes; here is an example:

> 'Which of the following 5 makes the best comparison?'
> 'Son is to Father as nephew is to . . . .'

Respondents would then be given the following choices: (a) Niece, (b) Cousin, (c) Uncle, (d) Mother, (e) Sister.

- Motivation and attitude measures are usually concerned with measuring particular beliefs towards something, such as work. So, for example, respondents would usually be asked to respond to an item such as 'I am satisfied with the work I do'.
- Neuropsychological tests are using measures of sensory, perceptual and motor performance used to assess different parts of psychophysiological activity and neurological functioning within the brain.

So, how are these different psychometric tests used? For example, in educational psychology, ability tests are used in the study of educational success and may be used as a tool in school placement, in detecting possible learning disabilities, or in tracking intellectual development. Personality tests are widely used in occupational psychology, particularly in job selection. What employers do is create their job criteria and then go some way in trying to match applicants to these criteria via personality testing. For example, if an employer wanted someone to sell a product, they would want that person to be outgoing. Therefore, the employer might administer an extraversion test to all applicants to see which of the candidates were more outgoing. Clearly, intelligence and attitude tests (particularly around motivation to work) are also used in occupational settings. Psychometric tests are used in clinical psychology as a way of diagnosing clinical conditions and distinguishing between clinical groups. For example, a clinical psychologist might compare their current treatment group on a measure against a general population sample, as this might provide a useful insight into how they should treat the clinical group. Equally they may use neuropsychological tests to assess consequences of medical illnesses or conditions. One example of such use might be among patients who have experienced brain damage; certain psychometric tests, particularly ability tests, might also be used to evaluate the extent of the damage and to later evaluate any improvement in a patient's condition.

Throughout the rest of the chapter we are going to introduce you to many of the aspects of psychometric testing by way of developing our own psychometric test.

Developing a psychometric test

For this exercise we're going to detail the development of a new measure of a concept, academic vindictiveness among students. As you might realise there are many approaches that students might take in their study. We've seen from Chapter 16 (The Application of Personality and Intelligence in Education and the Workplace), which details personality, intelligence and education, that conscientiousness is a good predictor of education performance. However, for the purpose of this chapter, we're wondering whether there might be another individual difference variable in the way students approach their academic work, namely academic vindictiveness.[1] This work was carried out in response to a finding that some students may make deliberate acts to sabotage others work, and/or feel angry towards someone or act vengefully when they feel somebody in their academic circle has wronged, misguided or surpassed them in some way (Crocker, Sommers and Luhtanen, 2002; Crocker and Luhtanen, 2003). Therefore, in the next few sections we will be using this construct of academic vindictiveness, and the development of a scale to measure academic vindictiveness to introduce you to a number of psychometric techniques.

Developing items for a psychometric test

Paul Kline (Kline, 1986), a United Kingdom psychometrician, points out that the secret to developing a very good scale is writing very good questions. For Kline, if you do not write good items then you will never develop a good psychometric scale. Therefore, for Kline, developing the items is a crucial part of the process. There are a number of considerations you need to make when writing items for a scale, and in the next section we will introduce you to these considerations while developing items for our own scale of academic vindictiveness.

The first thing to consider is to make the distinction between two different types of questions: open format or closed format questions. Open format are questions that asked for some written detail but have no determined set of responses, e.g. 'Tell us about the occasions when you have been academically vindictive'. Therefore, any answer can be given to these questions. These types of questions lead to more qualitative data because there are a large number of possible responses. These are the problems with open-format

[1]This is a construct that was developed by a colleague R. J. Lally and we are very grateful to her for allowing us to use some of her initial work on this construct in this chapter.

questions in a psychometric test. Given that they produce a large number of possible answers, with each potentially different, open format questions are also time-consuming for the researcher who has to analyse all the different answers.

As a consequence, in psychometric tests, researchers will tend to use closed-format questions. A closed format question is a question where there is a short question or statement followed by a number of options. See, for example, Exhibit 25.1.

Exhibit 25.1

Indicate the extent you disagree or agree with the following statement, as it applies to you.

1	I feel bitter towards those who do better than me on my course.	Disagree strongly	Disagree	Not certain	Agree	Agree strongly

The main aim of closed format questions is to very simply break down respondents answers into data that can be quantified into answers that give you the information you want to know (i.e. to what extent do people agree or disagree with the statement as it applies to them).

However, the first stage of the process is to create the items. Kline suggests that the first place to start is to write as many items as possible. Now, this could be a matter of simple writing the items yourself, but this can be very laborious and you may also make mistakes or miss out important aspects. Therefore, in terms of initial item writing, you could also use the following sources in writing items:

- **Theoretical literature** – Usually scales are not developed from an entirely new construct. There will always be some theoretical perspective which underpins the development of a new scale and the terms, phrases, and ideas that appear in the theoretical literature should be used as the basis for writing items.
- **Experts** – You could recruit experts in the area to suggest particular items for your scale. This will enhance the quality.
- **Colleagues** – Colleague(s) or a co-researcher(s) can help you write items because, at the very least, this will generate more items than you could singly produce.

After you have written the initial set of items you need to study them and rewrite them. Again, use experts or colleagues to examine the phrasing of each question and see if they can improve on them. It is also probably the case that you will have a large number of items.

It may then be a good idea to try to reduce the number of items you have written if you feel there are a huge number. It's difficult to determine what a huge number is, but in deciding on the final number you might like to keep a few points in mind:

- Kline suggests optimal length for any scale measuring one construct should be about 15 to 20 items. You may need many more items than this when initially constructing the scale to ensure you end up with a scale of an appropriate length.
- When you administer the scale in the first instance the general rule is that you should have a certain number of participants for each item of the administration; guidelines change, but at least 5 participants to 1 item is acceptable, but ideally most researchers aim for 10 participants to 1 item as a premium, with a minimum of 100 respondents. This is for statistical reasons because you need a good number of responses to ensure you are sure you have captured variation of responses across respondents. You might keep this in mind if you know how many participants you are likely to get to fill in your scale on a first occasion. For example, if you had 40 items, you would need 400 participants to answer the questionnaire. Therefore, if you were only likely to get 300 participants, you might wish to reduce your items, or lower the criteria to 5 to 1.
- Who is the scale likely to be used with? It may be that if the scale is designed for clinical settings, or with children, or where available administration time is short, you might actually look to reduce the number of items, even before you've developed the final scale.

It is difficult to say what the optimal number of items is. In most cases you need to make a judgement. One way to determine this is to get a group of participants, ideally experts, to rate the items in terms of potential effectiveness of measuring the construct. You could determine which items are considered the strongest and should be retained (i.e. those items rated highest by the participants) and which items are considered the weakest and could be excluded. This technique could be used to exclude items from the first administration of the scale if you felt you had too many items.

Writing items for a psychometric test

To illustrate the process and some of the techniques used in writing items we are going to talk about the development of a set of 23 items to measure vindictiveness. In developing these sets of items we followed many of the procedures outlined above. A few lecturers and students

wrote a list of 70 items. We then asked a group of five students to look at the items and suggest possible changes to the wording of the items. We then repeated the exercise with three lecturers.

For the purpose of the exercise we felt that 70 items were too many, so we decided to trim these down. We asked 5 students and 5 lecturers to rate the items in terms of their potential to measure academic vindictiveness, on a scale of 1 to 10 (1 = Not at all, 10 = Very much so). We then selected the top scoring items (those that scored an average rating score of 9 or over [though it is important to note this was an arbitrary criterion]. This amounted to 23 items, which we thought was a suitable number for the current exercise, and these are shown in Exhibit 25.2.

Exhibit 25.2

Potential Items for our Academic Vindictiveness Scale.

1. I can be spiteful to my friends if they get a better mark than me.
2. In the past I have falsely told other students the wrong exam date, but only when they were too lazy to find out themselves, and only when I was in a bad mood, so they would miss the exam.
3. I would hate the student who got the best mark in one of my classes.
4. If I had the opportunity, resources and ability to change other students' exam grades so that mine were the best, I would do it.
5. I find myself wishing bad things on people that do better than me academically.
6. I always tell people I am happy for them when they do better than me in exams.
7. I have thought about spoiling someone's work because it is better than mine.
8. I wish bad things on people because they are smarter than me.
9. When other students on my course are praised for their excellent work, it makes me want to wish something bad on that person.
10. If my friend got a better mark than me, even if it was just 1 per cent, I would consider tampering with his/her work in some way.
11. I have mean thoughts towards people who score better than me in exams.
12. I resent people on my course who excel in their studies.
13. I feel bitter towards those who do better than me on my course.
14. I have lied to people on my course to try and hinder their progress so I can get the better mark.
15. If I came second best in a piece of work, because my friend got the top mark, I would still be happy.
16. I am a really bad person because I am academically vindictive.
17. I seek revenge on people who get better grades than me and take away my chance of success.
18. If I was Vice-Chancellor of an University I would ensure that students were not academically vindictive.
19. I have thought about spoiling someone's work because it is better than mine.
20. I would consider doing something nasty to somebody who threatened my chances of academic success.
21. My academic vindictiveness is a result of a mental illness.
22. When I find out I haven't got an excellent mark for a piece of work I get very upset with myself.
23. I work very hard so I am able to do the best I can at all academic activities.

Now, we have deliberately made some mistakes or errors in some of these items so we can demonstrate some of the skills in item writing: clarity, leading, embarrassing, hypothetical, and reverse questions.

Clarity of questions

You have to ensure that there is clarity to the wording of your questions. Good practice suggests that questions must be clear, short and unambiguous. Look at the following questions, 4 and 10.

4. If I had the opportunity, resources and ability to change other student's exam grades so that mine was the best, I would do it.
10. If my friend got a better mark than me, even if it was just 1 per cent, I would consider tampering with his/her work in some way.

The main aim of writing a good psychometric test question is to make sure that the questions will not mean different things to different respondents. This is very important because what it means is that if your questions are ambiguous

(the meaning can be interpreted differently) then your participants will, in fact, be answering different questions. This will muddy your results because you will never be sure what interpretation respondents have been answering. One of the main culprits of this is qualifying statements or trying too hard to capture all aspects of the situation. Question 4 is an example of an item trying to capture all the prerequisites that might have to be in place that possibly underlie academic vindictiveness (opportunity, resources and ability). This means that respondents might concentrate on aspects of the question that focus on the opportunity, resources and ability before they were academically vindictive, rather than whether they have the tendency to be academically vindictive, therefore making the question ambiguous. Question 10 seeks to qualify the possible extremity of academic vindictiveness by highlighting issues about academic vindictiveness, even if the gap in scores was less than 1 per cent. Each of these questions can be simplified as shown in Exhibit 25.3.

Exhibit 25.3

4 If I had the opportunity to change other student's exam grades so that mine was the best, I would do it.

10 If my friend got a better mark than me, I would consider tampering with his/her work in some way.

These questions are shorter and the meaning is much more clear. They may not seek to be as exact as the previous versions, through the use of qualifying statements, but they are unlikely to be ambiguous to respondents.

Leading questions

Leading questions are questions that try to steer the respondent to a particular answer, or in the direction of a particular answer.

However, leading questions can arise quite undeliberately and can occur to the exact phrasing of the question. A brilliant example of the use of leading questions was demonstrated in one episode of the 1986 *Yes, Prime Minister* BBC series: *Yes, Prime Minister* is a fictional BBC comedy series set in the Prime Minister's office in 10 Downing Street and follows the Prime Ministerial career of Jim Hacker and struggles to formulate and enact legislation or effect departmental changes opposed by the will of the British Civil Service. Sir Humphrey Appleby, and his Principal Private Secretary Bernard Woolley. The interchange between Sir Humphrey and Bernard Woolley demonstrates how surveys can reach opposite conclusions about the introduction of national service (compulsory military *service*) for young people through the use of leading questions (see Exhibit 25.4).

Exhibit 25.4

Survey one

Sir Humphrey Appleby: Mr Woolley, are you worried about the rise in crime among teenagers?

Bernard Woolley: Yes.

Sir Humphrey Appleby: Do you think there is lack of discipline and vigorous training in our Comprehensive Schools?

Bernard Woolley: Yes.

Sir Humphrey Appleby: Do you think young people welcome some structure and leadership in their lives?

Bernard Woolley: Yes.

Sir Humphrey Appleby: Do they respond to a challenge?

Bernard Woolley: Yes.

Sir Humphrey Appleby: Might you be in favour of reintroducing National Service?

Bernard Woolley: Er, I might be.

Sir Humphrey Appleby: Yes or no?

Bernard Woolley: Yes.

Sir Humphrey Appleby: Of course, after all you've said you can't say no to that. On the other hand, the surveys can reach opposite conclusions.

Survey two

Sir Humphrey Appleby: Mr Woolley, are you worried about the danger of war?

Bernard Woolley: Yes.

Sir Humphrey Appleby: Are you unhappy about the growth of armaments?

Bernard Woolley: Yes.
Sir Humphrey Appleby: Do you think there's a danger in giving young people guns and teaching them how to kill?
Bernard Woolley: Yes.
Sir Humphrey Appleby: Do you think it's wrong to force people to take arms against their will?
Bernard Woolley: Yes.
Sir Humphrey Appleby: Would you oppose the reintroduction of conscription?
Bernard Woolley: Yes.

As you can see Sir Humphrey has been able to get Bernard Woolley to reach two opposing conclusions by using leading questions that Bernard Woolley will tend to agree with.

Our question 16 is a leading question 'I am a really bad person because I am academically vindictive'. It is asking respondents to make a value judgement that they would find it difficult to disagree with, and therefore they would be led to answer because they feel they would be saying something about themselves and place themselves in a bad light. We would suggest removing this question.

Perhaps our question 2 is a less obvious example of a leading question, see Exhibit 25.5.

Exhibit 25.5

2 In the past I have falsely told other students the wrong exam date, but only when they were too lazy to find out themselves, and only when I was in a bad mood, so they would miss the exam.

Question 2, seeks to qualify the item asking about academic vindictiveness by closely defining the situation around the other student being lazy and whether the person was in a bad mood. This is clearly a situation in which academic vindictiveness may occur, but also seems to lead the respondent in a particular direction by potentially excusing the behaviour and almost suggesting that it could be acceptable under circumstances. Regardless of whether it encourages people to 'own up' it is leading the respondent by suggesting they could be justified to make the statement. We would suggest, in terms of directly assessing academic vindictiveness, a better item might be as shown in Exhibit 25.6.

Exhibit 25.6

2 In the past I have falsely told other students the wrong exam date, so they would miss the exam.

Embarrassing questions

Generally, questions dealing with personal matters should be avoided. This is because this may make your respondent feel embarrassed or uncomfortable, and doing research that causes this type of feeling in participants is not good practice, indeed it is frowned upon. Equally, it is not good for the researcher because it may lead the participant to give incorrect or misleading information, or fail to complete the rest of the questionnaire. So, great care should be taken when asking personal or potentially embarrassing questions and you should spend a lot of time thinking about how best to ask these questions. Our question 21 ('My academic vindictiveness is a result of a mental illness') is a potentially embarrassing question and we should remove this item from our list.

Questions dealing with personal matters should be avoided, unless absolutely necessary, as they have the potential to cause embarrassment.
Source: Image Source/Rex Features

Hypothetical questions

Hypothetical questions are questions that place the individual in a situation that they may never experience and ask them for their opinion on something. So, for example, you might ask 'If you were Prime Minister of the country, what would you do about psychology lecturers'. These types of questions might produce colourful answers, but are considered bad research practice because answers will be in response to a situation the person may never have considered rather than their real view or feelings about something. Our question 18 ('If I was Vice-Chancellor of a University I would ensure that students were not academically vindictive') is a hypothetical question and therefore we should remove it from our list.

Questions with reverse wording

Researchers will often write items with some questions with *reverse wording*. This is usually done to force the person taking the psychometric test to read it carefully and not just to respond to all the items in the same way. Also, it is good way to check for people who haven't taken answering the questionnaire seriously, as it will show up with contradictory answers. Ideally, you should try to maximise the number of reverse worded items in your scale, but we have just included reverse wording for two items, item 6 and item 15.

Therefore, taking all these changes into account, we have a final revised scale comprising 20 items as shown in Exhibit 25.7.

Exhibit 25.7

1 I can be spiteful to my friends if they get a better mark than me.
2 I find myself wishing bad things on people that do better than me academically.
3 I would hate the student who got the best mark in one of my classes.
4 I feel bitter towards those who do better than me on my course.
5 In the past I have falsely told other students the wrong exam date, so they would miss the exam.
6 I always tell people I am happy for them when they do better than me in exams.
7 I have thought about spoiling someone's work because it is better than mine.
8 I seek revenge on people who got better grades than me and take away my chances of success.
9 When other students on my course are praised for their excellent work, it makes me want to wish something bad on that person.
10 If my friend got a better mark than me, I would consider tampering with his/her work in some way.
11 I have mean thoughts towards people who score better than me in exams.
12 I resent people on my course who excel in their studies.
13 If I had the opportunity to change other students' exam grades so that mine were the best, I would do it.
14 I have lied to people on my course to try and hinder their progress so I can get the better mark.
15 If I came second best in a piece of work, because my friend got the top mark, I would still be happy.
16 I wish bad things on people because they are smarter than me.
17 I do not like to help others with their work as it might result in them getting a better mark.
18 I would consider doing something nasty to somebody who threatened my chances of academic success.
19 When I find out I haven't got an excellent mark for a piece of work I get very upset with myself.
20 I work very hard so I am able to do the best I can at all academic activities.

Stop and think

? ! ?

Cultural considerations

There is other good practice in writing good questions. Try to consider language or culture and make sure that all questions can be understood by all people. Remember, people sometimes have a poor reading age so try to make the questions as simple as possible. Also some questions may seem patronising to you, but people who have a high reading age will still be able to understand the question. Try to avoid jargon or technical language, particularly abbreviations.

Response formats

Another important area to consider is the response format of your scale. All closed format questions give a series of choices, and there is even good practice in terms of response choices to use. Of course, the response formats generally vary.

There are a number of different formats that can be used. One of the response formats used with a lot of traditional personality tests is a 'yes–no' format or a 'true–false' format. For example, the Eyenck Personality Questionnaire (Eysenck and Eyseck, 1975) uses this sort of format (see Exhibit 25.8).

Exhibit 25.8

1 Does your mood often go up and down?	Yes	No
2 Do you often feel 'fed-up'?	Yes	No
3 Do you suffer from nerves?	Yes	No

Other scales measure the frequency of behaviour (i.e. how often it occurs). For example, the COPE scale (Carver, Scheier and Weintraub, 1989), that measures individual's various coping attempts to deal with stress, uses the response format shown in Exhibit 25.9 to measure the frequency to which people engage in various coping behaviours.

Exhibit 25.9

I try to grow as a person as a result of the experience.	I usually don't do this at all	I usually do this a little bit	I usually do this a medium amount	I usually do this a lot
I turn to work or other substitute activities to take my mind off things.	I usually don't do this at all	I usually do this a little bit	I usually do this a medium amount	I usually do this a lot
I get upset and let my emotions out.	I usually don't do this at all	I usually do this a little bit	I usually do this a medium amount	I usually do this a lot

A common feature of many scales is the 'Strongly Agree' to 'Strongly Disagree' format. Traditionally, these were used with attitude scales, used to measure the extent of agreement with different attitudinal statements. However, you will see the 'agree–disagree' response format used in many different scales. For example, the Life Orientation Test-Revised (Scheier, Carver and Bridges, 1994), which is used to measure optimism, uses a five-point 'Strongly Agree' to 'Strongly Disagree' scale to measure respondent's degree of agreement with statements in terms of how the statements describe them, as shown in Exhibit 25.10.

Exhibit 25.10

In uncertain times, I usually expect the best.	Strongly Disagree	Disagree	Not certain	Agree	Strongly Agree
Overall, I expect more good things to happen to me than bad.	Strongly Disagree	Disagree	Not certain	Agree	Strongly Agree

Though sometimes some scales will ask respondents to indicate directly how much the statement describes them. For example, in the assessment of dispositional embarrassment (Kelly and Jones, 1997) respondents are asked to indicate to what extent the behaviour described in the statement is like them (see Exhibit 25.11).

Exhibit 25.11

	Not at all like me						Very much like me
1 I feel unsure of myself.	1	2	3	4	5	6	7
2 I don't feel uncomfortable in public unless my clothing, hair, etc. are just right.	1	2	3	4	5	6	7

Finally, you will also see response formats that try to assess the extent of certain feelings or behaviours, and therefore, the responses will assess the extent that the respondent feels about something. So, for example, the PANAS scale (Watson, Clark and Tellegen, 1988) which measures positive and negative affect, uses a scale that indicates the extent the participant feels about a particular emotion or feeling (see Exhibit 25.12).

Exhibit 25.12

Interested	Very slightly	A little	Moderately	Quite a bit	Extremely
Irritable	Very slightly	A little	Moderately	Quite a bit	Extremely
Distressed	**Very slightly**	**A little**	**Moderately**	**Quite a bit**	**Extremely**

There are fewer and fewer hard and fast rules about response formats these days. However, the main point is that your response format must make sense in terms of the questions you are asking. Therefore discuss with colleagues and test carefully the response format that you intend to use. Also, a general guideline is to use simple rating scales or lists of choice, and where possible, minimise the number of choices. Five choices are thought to be the ideal number.

Instructions

Finally, the instructions that precede the scale are crucial. Usually there may be no need for a great number of instructions, and they be rather simple. For example, the instructions for Eysenck and Eysencks' EPQ scale are as shown in Exhibit 25.13.

Exhibit 25.13

INSTRUCTIONS: Please answer each question by putting a circle around the 'YES' or 'NO' following the question. There are no right or wrong answers, and no trick questions. Work quickly, and do not think too long about the exact meaning of the questions.

However, you might look for more general traits reflecting more typical behaviours or attitudes. For example, for Kelly and Jones' measure of dispositional embarrassment, the instructions are as shown in Exhibit 25.14.

Exhibit 25.14

We are interested in people's personality attributes. Listed below are a variety of statements. Please read each statement carefully and indicate to the left of each item the extent to which you feel it applies to you using the following scale.

However, you may want to specify a particular time period. For example, in completing Watson *et al.*'s Positive and Negative Affect Scales, respondents are given the following instructions as shown in Exhibit 25.15.

Exhibit 25.15

This scale consists of a number of words that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent you feel this way right now, that is, at the present moment. Use the following scale to record your answers.

Finally, you might want the respondents to think very carefully about their responses. For example, you might want them to think about a specific instance, or a typical set of responses in a particular circumstance. For example, Carver *et al.*'s COPE scale, which measures reactions to stress, is preceded by the instructions shown in Exhibit 25.16.

Exhibit 25.16

We are interested in how people respond when they confront difficult or stressful events in their lives. There are lots of ways to try to deal with stress. This questionnaire asks you to indicate what you generally do and feel, when you experience stressful events. Obviously, different events bring out somewhat different responses, but think about what you usually do when you are under a lot of stress.

Please try to respond to each item separately in your mind from each other item. Choose your answers thoughtfully, and make your answers as true FOR YOU as you can. Please answer every item. There are no 'right' or 'wrong' answers, so choose the most accurate answer for YOU – not what you think 'most people' would say or do. Indicate what YOU usually do when YOU experience a stressful event.

The main thing is that the researcher thinks carefully about the instructions, because this can be used not only to help to make easier the administration of the question but also to direct the respondent to look at the questions in a particular way, if so required.

Collecting the data

Following the advice from the following section we now have enough information for our scale. Our proposed scale is now shown in Exhibit 25.17.

Exhibit 25.17

Instructions: The following statements refer to your *own* beliefs and feelings about your own thoughts, feelings and behaviour. Read each statement and respond by circling the number that best represents your agreement with each statement.

[1 = Strongly disagree, 2 = Disagree, 3 = Not certain, 4 = Agree, 5 = Strongly agree]

1	I can be spiteful to my friends if they get a better mark than me.	1	2	3	4	5
2	In the past I have falsely told other students the wrong exam date, so they would miss the exam.	1	2	3	4	5
3	I would hate the student who got the best mark in own of my classes.	1	2	3	4	5
4	If I had the opportunity to change other students' exam grades so that mine was the best, I would do it.	1	2	3	4	5
5	I find myself wishing bad things on people that do better than me academically.	1	2	3	4	5
6	I always tell people I am happy for them when they do better than me in exams.*	1	2	3	4	5
7	I have thought about spoiling someone's work because it is better than mine.	1	2	3	4	5
8	I wish bad things on people because they are smarter than me.	1	2	3	4	5
9	When other students on my course are praised for their excellent work, it makes me want to wish something bad on that person.	1	2	3	4	5
10	If my friend got a better mark than me, I would consider tampering with his/her work in some way.	1	2	3	4	5
11	I have mean thoughts towards people who score better than me on exams.	1	2	3	4	5
12	I resent people on my course who excel in their studies.	1	2	3	4	5
13	I feel bitter towards those who do better than me on my course.	1	2	3	4	5
14	I have lied to people on my course to try and hinder their progress so I can get the better mark.	1	2	3	4	5
15	If I came second best in a piece of work, because my friend got the top mark, I would still be happy.*	1	2	3	4	5

16	I seek revenge on people who get better grades than me and take away my chances of success.	1	2	3	4	5
17	I do not like to help others with their work as it might result in them getting a better mark.	1	2	3	4	5
18	I would consider doing something nasty to somebody who threatened my chances of academic success.	1	2	3	4	5
19	When I find out I haven't got an excellent mark for a piece of work I get very upset with myself.	1	2	3	4	5
20	I work very hard so I am able to do the best I can at all academic activities.	1	2	3	4	5

The next step is to collect data to test these new items and assess whether you've got a good measure. Therefore, we need to administer the scale to some participants. Guidelines suggest that you should try to get a minimum of 5 respondents to every item, ideally the number of respondents should be 10 participants for every 1 item. Of course, as a student you may not have access to that number of participants, but you should try to get at least 2 or 3 respondents for each item.

We administered our scale to 402 students, 161 males and 241 females, aged from 18 to 21 years, meaning we had a ratio of 20 to 1 in terms of participants to items. Having collected the data we can now use this data to examine the scale.

We have included a copy of this data as part of the online resources with this book. Go to **pearsoneducation.co.uk/maltby** if you want to download the data.

Reliability

In psychometric testing, there are two forms of reliability: internal reliability and reliability over time (test–retest reliability). You may wish to read an extended version of some of the theory that lies behind reliability statistics in the 'Stop and think: Reliability: the role of error' box.

Stop and think

? ! ?

Reliability: the role of error

Reliability in psychometric testing is a response to error in measurement. Error refers to a specific issue in research and that is, when you're measuring anything in research, it is almost certain that your measurement will contain error. Say, for example, I ask you the question, are you happy? I give you the option of 'Yes' or 'No'. Now what I'm hoping to do is measure your happiness. However, there are possible sources of error in this measurement. For example, you may say 'Yes', therefore you are happy. But you may not be totally happy, you may just be more happy than you are unhappy, but my measure just determines that you are happy. Therefore there is possible error in my measurement because I've not exactly assessed the correct level of happiness, so my assessment of your happiness is not wrong (you've said you're happy), but it has a degree of error to it. There are other potential sources of error in measurement. Perhaps if I asked you two questions; are you happy in 'Work' and 'Life in general' to determine your happiness. If you answered 'Yes' to both then you are happy. But it is perfectly possible that you are not happy in other specific areas of your life, such as a relationship. Therefore again there is potential for error here because our measure of happiness missed out a question on relationships which would have changed our assessment of your overall happiness. Also if I asked you whether you were happy on a Monday, would you give the same answer on the following Friday?

There are many sources of possible error in measurement. It is important to note that these errors are mostly unknown quantities and often not measurable. For example, it is impossible to know what real happiness is. Therefore asking people are they happy and determining on that whether they are happy or not has a huge possible amount of error, because the extent and depth of happiness is probably unknown and probably would make your head hurt just thinking about how to measure it. However, rather than simply giving up, researchers persevere. As a researcher it is almost impossible to eradicate all possible sources of error from research, but what researchers do is try to guard against possible error so they can establish confidence in their work. Reliability statistics are used in psychometric testing to assess the extent to which a psychometric test is free from error and provide confidence and evidence for its usefulness.

Internal reliability (internal consistency)

Internal reliability (or consistency) refers to whether all the aspects of the psychometric test are generally working together to measure the same thing. Therefore, we would expect to find that all these aspects would be positively correlated with each other. Commonly, these aspects would be a number of questions in a scale. Therefore, all the individual questions on our academic vindictiveness scale should correlate, suggesting they go together to form a single construct of academic vindictiveness.

A common statistical technique that you will see in the research literature, that is used to assess the internal reliability, is Cronbach's alpha (Cronbach, 1951). Cronbach's alpha, then, is used to assess the level of internal reliability that a set of items has. The figure that is produced to assess this level can range between –1 and +1, and its symbol is α. Usually, a Cronbach's alpha of +0.7 or above is seen as an acceptable level of internal reliability, although in some circumstances a much higher level is desired (i.e. $\alpha > 0.8$) when a more exact measurement is wanted or required, i.e. in clinical assessments of patients.

You can work out the internal reliability of our items using SPSS for Windows. Load up the data, and in the data screen Window, Click on Analyse, Scale and select reliability analysis. Transfer the items (AV1 to AV20) into the items box (see Figure 25.1). Please note for this analysis you need to make sure you have recoded any reverse items. We have recoded the scores (using the RECODE statement in SPSS) for the reverse worded items, 6 and 16.

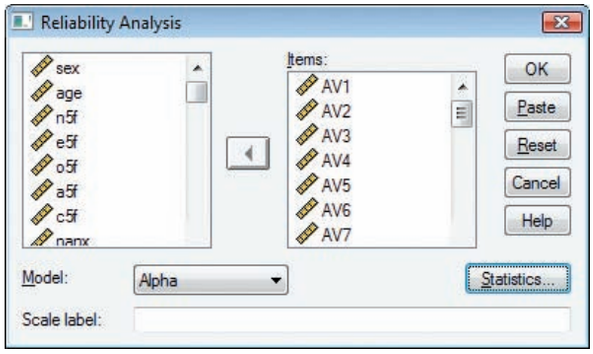


Figure 25.1 Reliability analysis window.

Then press OK. You should get the output shown in Table 25.1.

Table 25.1 Reliability analysis output.

Reliability statistics

Cronbach's alpha	*N* of items
0.866	20

Stop and think

? ! ?

Other tests of internal reliability

There are other tests of internal reliability that are used less often in psychometric testing, but which, at some point, may prove useful to you.

In psychometrics, the **Kuder–Richardson Formula 20** (KR-20) is a measure of internal reliability for measures with dichotomous choices (i.e. 2 choices, Yes/No – Agree/Disagree). Many instruments have response formats that are dichotomous, and technically you should not perform a Cronbach's alpha on this. That said, many researchers do use Cronbach's alpha with measures with dichotomous choices. However, it may be useful for you to have the alternative. Values can range from 0.00 to 1.00 with higher values indicating a better level of internal reliability. The optimum level for internal reliability is within a KR-20 of 0.80 to 0.85 range. The Kuder-Richardson Formula 20 (KR_{20}), for example, calculates a reliability coefficient based on the number of test items:

$$r = \frac{k}{(k-1)}\left(1 - \frac{\Sigma pq}{s^2}\right)$$

Where

k = is the number of test items

p = the proportion of the responses to an item that are correct or have been answered in one direction (i.e. yes or agree); i.e. the number of correct (or 'yes'/'agree') answers out of the total number of responses.

q = the proportion of responses that are incorrect to an item that are incorrect or have been answered in the other direction (i.e. no or disagree); i.e. the number of incorrect (or 'no'/'disagree') answers out of the total number of responses.

s^2 = the variance, or the standard deviation squared.

Another form of internal reliability is *split-half reliability*. Here the research splits the items into two halves. This split might be made based on odd versus even numbered items, randomly selecting items for each half, or the first half versus the second half of the test. The researcher then correlates the total scores for each half. A common rule of thumb is 0.80 or high for adequate reliability, though practice does vary.

Here the Cronbach's alpha is $\alpha = 0.866$ (or $\alpha = 0.87$, as it is normally rounded up to two decimal places). The current Cronbach's alpha is good, above $\alpha = 0.7$, and therefore is of acceptable internal reliability.

Using internal reliability to select items

However, we can look a little closer at our scale to see if the internal reliability of the scale can be improved. This is particularly useful if the reliability of your items has fallen below the criteria of $\alpha = 0.7$. You should routinely do this anyway, because the procedures shown in Figure 25.2 allow you to identify items that are performing poorly.

To perform this analysis in SPSS for Windows repeat the above analysis for the reliability, but before pressing the OK button, press the statistics button and tick the three boxes in the **Descriptives for** section (Item, Scale and Scale if item deleted (see Figure 25.2)). Then press Continue and then OK. You should then get an output that looks like that shown in Table 25.2.

Table 25.2 gives two pieces of information that are important to us, Corrected Item-Total Correlation (column 4) and the Cronbach's Alpha if (the) Item (was) Deleted (column 5). The Corrected Item-Total Correlation tells us how much each item is related to the overall score, and the Cronbach's Alpha if Item Deleted tells us what the Cronbach's alpha of the scale would be if it was actually deleted. With

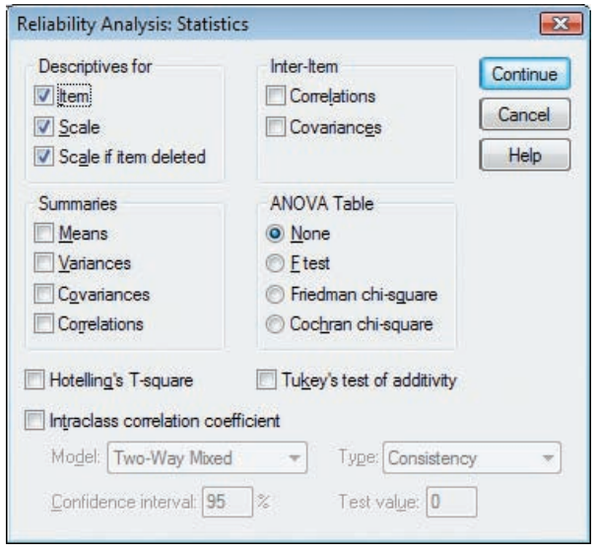


Figure 25.2 Reliability analysis: statistics window.

Table 25.2 Reliability statistics output.

Item-total statistics

	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
AV1	46.8234	133.158	0.509	0.858
AV2	46.9353	133.632	0.611	0.854
AV3	47.3358	136.797	0.540	0.857
AV4	46.4453	132.402	0.535	0.857
AV5	46.9776	132.655	0.608	0.854
AV6	46.8632	134.158	0.532	0.857
AV7	46.7836	135.956	0.426	0.861
AV8	46.9602	133.320	0.546	0.856
AV9	47.1169	136.921	0.453	0.860
AV10	46.9080	136.747	0.435	0.860
AV11	46.8607	133.796	0.604	0.855
AV12	47.4303	136.814	0.595	0.856
AV13	46.8159	135.288	0.520	0.857
AV14	46.6418	133.831	0.516	0.857
AV15	46.7612	133.743	0.512	0.857
AV16	46.6741	132.978	0.644	0.853
AV17	46.8682	136.070	0.552	0.857
AV18	46.7264	132.947	0.508	0.858
AV19	45.4179	151.730	−0.109	0.880
AV20	45.4353	152.885	−0.146	0.882

Table 25.3 Reliability statistics output.

Reliability statistics

Cronbach's alpha	*N* of items
0.882	19

Item-total statistics

	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
AV1	43.0597	135.847	0.512	0.876
AV2	43.1716	136.402	0.611	0.873
AV3	43.5721	139.417	0.548	0.875
AV4	42.6816	135.145	0.535	0.875
AV5	43.2139	135.101	0.621	0.872
AV6	43.0995	136.748	0.539	0.875
AV7	43.0199	138.364	0.440	0.878
AV8	43.1965	135.819	0.557	0.874
AV9	43.3532	139.446	0.464	0.877
AV10	43.1443	139.371	0.442	0.878
AV11	43.0970	136.626	0.602	0.873
AV12	43.6667	139.455	0.604	0.874
AV13	43.0522	137.980	0.524	0.875
AV14	42.8781	136.471	0.521	0.875
AV15	42.9975	136.421	0.516	0.876
AV16	42.9104	135.658	0.648	0.872
AV17	43.1045	138.852	0.553	0.875
AV18	42.9627	135.742	0.507	0.876
AV19	41.6542	155.309	−0.130	0.897

these statistics we can see if any items are not contributing to the scale.

If we look at the Corrected Item-Total Correlation good items should correlate above 0.3 with the total score, and not be below 0.20 (Kline, 1986). We can see here that items AV19 and AV20 are below this criteria (AV19 = –0.109; AV20 = –0.146). The Cronbach's Alpha if Item Deleted column also suggests that for both these items the current alpha of $\alpha = 0.87$ would be improved, to $\alpha = 0.880$ (0.88) for AV19 and $\alpha = 0.882$ (0.88) for AV20. These statistics suggest the scale's Cronbach's alpha would be improved by removing these two items.

What you do in these circumstances is remove one item at a time, with the worst performing item (i.e. lowest correlation and most improved Cronbach's alpha) being the one that is removed first, and then seeing what effect it has on the results. You continue this until there is no more improvement.

In our example, the worse performing item is Item AV20. Therefore we recompute the Cronbach's alpha with all the items except AV20. We then get an output that looks like Table 25.3.

We can see that the Cronbach's alpha coefficient has improved to $\alpha = 0.88$ (from 0.87) with the removal of AV20. However, if we look at the bottom table we can see that item AV19 shares a low negative correlation (–0.130) with the overall score and removal of the item will improve the Cronbach's alpha to $\alpha = 0.897$ (0.90 to two decimal places). Therefore, repeating the procedure and removing AV19 produces Table 25.4.

You can see now that all the corrected item-total correlations are satisfactory (all above 0.5) and our alpha coefficient of 0.90 can't really be improved upon if we look at the figures in the Cronbach's Alpha if Item Deleted column. Consequently we would use the first 18 items for our Academic Vindictiveness scale.

This type of analysis is useful for developing a scale, particularly when you first run the alpha coefficient, and the alpha coefficient falls below a satisfactory criteria. It may be that by removing some items in this manner, one item at a time, you can improve the internal reliability of the scale.

Table 25.4 Reliability statistics output.

Reliability statistics

Cronbach's alpha	*N* of items
0.897	18

Item-total statistics

	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
AV1	39.2786	138.012	0.516	0.892
AV2	39.3905	138.683	0.612	0.889
AV3	39.7910	141.617	0.554	0.891
AV4	38.9005	137.691	0.526	0.892
AV5	39.4328	137.104	0.632	0.889
AV6	39.3184	138.866	0.546	0.891
AV7	39.2388	140.696	0.440	0.895
AV8	39.4154	137.994	0.561	0.891
AV9	39.5721	141.517	0.474	0.893
AV10	39.3632	141.504	0.449	0.894
AV11	39.3159	138.895	0.603	0.890
AV12	39.8856	141.478	0.618	0.890
AV13	39.2711	139.974	0.536	0.892
AV14	39.0970	138.766	0.521	0.892
AV15	39.2164	138.709	0.517	0.892
AV16	39.1294	137.983	0.646	0.888
AV17	39.3234	141.162	0.553	0.891
AV18	39.1816	138.049	0.507	0.893

Stop and think

? ! ?

Mean scores and items

Another way of checking your items to see if the items are appropriate for your test is to look at the mean scores. This will give us some idea of the average scores. Why is this useful? Well say, for example, one of the mean scores for one of the items was particularly high. For example, for an item scored from 1 (Disagree strongly) to 5 (Agree strongly), a high mean (e.g. Mean = 4.8) would mean that almost all people were just simply Agreeing strongly with an item. This would mean that the item produced little variance in responses, and would almost be redundant for your scale because it would not differentiate between respondents, i.e. almost all people answering that item would answer it in the same way.

We have in the table below listed the Mean and Standard Deviations of all our 18 items selected so far, as well as the minimum and maximum scores obtained from our sample for each item. As you can see we haven't got any extreme means (i.e. very high or low means), with most of the means being above 2. However, two items are worth further consideration; 'I would hate the student who got the best mark in own of my classes' (item 3) and 'I resent people on my course who excel in their studies' (item 12). Clearly the mean scores of these items are a little lower than the rest of the items and it is also worth noting that the maximum score obtained for item 12 is 4 (rather than 5) meaning no-one Strongly agreed with this item. The lower mean scores of these items might be due to the phrasing of the items. For item 3, 'hate' is a rather strong word to use and that some people might not want to agree with an item that includes such a strong term. Similarly, for item 12, people might be concerned with using the phrase that 'resenting people who excel in studies' and that might be problematic if you yourself excel in your studies because that may mean that you resent yourself. These reasons are speculative. There is no evidence that these items are problematic, and may represent an accurate response to these statements. However, it would always be worth checking in

future analyses how these two items perform, because you may pick up other clues that these items are not performing as well as other items and come to conclusion that they could be dropped. Or indeed it might be worth considering rewording the item for future administrations. For example, with item 12 the clarity of the item might be improved by changing it to 'I resent *other* people on my course who excel in their studies'.

	Item	Minimum	Maximum	Mean	Std deviation
1	I can be spiteful to my friends if they get a better mark than me.	1.00	5.00	2.3756	1.28901
2	In the past I have falsely told other students the wrong exam date, so they would miss the exam.	1.00	5.00	2.2637	1.07343
3	I would hate the student who got the best mark in one of my classes.	1.00	5.00	1.8632	.96765
4	If I had the opportunity to change other student's exam grades so that mine were the best, I would do it.	1.00	5.00	2.7537	1.29127
5	I find myself wishing bad things on people that do better than me academically.	1.00	5.00	2.2214	1.14236
6	I always tell people I am happy for them when they do better than me in exams.*	1.00	5.00	2.3358	1.17078
7	I have thought about spoiling someone's work because it is better than mine.	1.00	5.00	2.4154	1.25106
8	I wish bad things on people because they are smarter than me.	1.00	5.00	2.2388	1.20377
9	When other students on my course are praised for their excellent work, it makes me want to wish something bad on that person.	1.00	5.00	2.0821	1.11249
10	If my friend got a better mark than me, I would consider tampering with his/her work in some way.	1.00	5.00	2.2910	1.16575
11	I have mean thoughts towards people who score better than me in exams.	1.00	5.00	2.3383	1.07336
12	I resent people on my course who excel in their studies.	1.00	4.00	1.7687	.88699
13	I feel bitter towards those who do better than me on my course.	1.00	5.00	2.3831	1.11105
14	I have lied to people on my course to try and hinder their progress so I can get the better mark.	1.00	5.00	2.5572	1.22442
15	If I came second best in an piece of work, because my friend got the top mark, I would still be happy.*	1.00	5.00	2.4378	1.23836
16	I seek revenge on people who get better grades than me and take away my chances of success.	1.00	5.00	2.5249	1.06667
17	I do not like to help others with their work as it might result in them getting a better mark.	1.00	5.00	2.3308	1.00000
18	I would consider doing something nasty to somebody who threatened my chances of academic success.	1.00	5.00	2.4726	1.30623

Test–retest reliability (reliability over time)

Test–retest reliability assesses reliability over time. Researchers interested in constructs that are concerned with individuals being relatively *consistent* in their attitudes and behaviours over time are interested in test–retest reliability. In the personality, intelligence and individual differences literature, which is a literature interested in traits, you will see reference to the stability of tests over time.

Say, for example, 402 respondents completed the newly developed academic vindictiveness questionnaire. We might be interested in finding out whether the test measured similar levels of academic vindictiveness in the respondents at another time. Researchers typically test people either a week, two weeks or a month apart from the first administration, though some researchers will also produce six-month, one-year and two-year intervals between administrations. The ability of the academic vindictiveness test to find similar levels of academic vindictiveness across the 100 respondents would provide evidence of its stability, and therefore its test–retest reliability. A correlation statistic (i.e., the researcher would hope that there is a significant positive correlation between the two-test administrations) is the most often used indicator of test–retest reliability and normally a value of the correlation of $r =$ or > 0.7 or above is considered as satisfactory.

Stop and think

? ! ?

What is a correlation?

As a test, the correlation coefficient can take values ranging from +1.00 through 0.00 to −1.00.

- A correlation of +1.00 would be a 'perfect' positive relationship.
- A correlation of 0.00 would be no relationship (no single straight line can sum up the almost random distribution of points).
- A correlation of −1.00 would be a 'perfect' negative relationship.

Commonly then, correlation statistics range from +1 to −1, and the symbol of a correlation is *r*. Therefore, researchers will report the direction of correlation coefficients between variables. For example, the relationship between neuroticism (anxious and worrying personality traits) and depression would be expected to fall within the 0.00 to +1.00 range, while the relationship between extraversion (outgoing, optimistic personality traits) and depression would be expected to fall within the 0.00 to −1.00 range.

However, it is important to remember that the reporting of correlation statistics doesn't stop there. There are two ways of interpreting the strength of the correlation. The first is the significance level. You remember that a lot of statistics involves interpreting whether a statistical test result is significant at either the 0.05 or 0.01 level. Therefore, commonly, researchers report whether the correlation is significant, be it a positive or negative relationship.

However, it is also necessary to highlight the size of the correlation (the *r* statistic). Researchers often do this to consider the weight of their findings, and this is also known as effect size (Cohen, 1988). A correlation statistic of $r = 0.1$ and below is viewed as small, $r = 0.3$ as medium (or moderate) and $r = 0.5$ as large. These are used as indicators of the relative importance of findings. Therefore, if a researcher has predicted that there will be a relationship between two variables, a positive correlation of 0.5 would be a more important finding than a correlation of 0.2.

Table 25.5 Correlation output.

Correlations

		ACADEMVIND	ACADEMVIND2
ACADEMVIND	Pearson Correlation	1	0.764(*)
	Sig. (2-tailed)		0.000
	N	402	402
ACADEMVIND2	Pearson Correlation	0.764(*)	1
	Sig. (2-tailed)	0.000	
	N	402	402

*Correlation is significant at the 0.01 level (2-tailed).

For academic vindictiveness we would assume this is a consistent trait and we could test our new academic vindictiveness scale for the scale's test–retest reliability. We asked all our respondents to fill in the scale again four weeks later. This data is provided in the data set, and by performing a Pearson Product moment correlation in SPSS (Analyse, Correlate, Bivariate and input our two variables ACADEMVIND [Time 1] and ACADEMVIND2 [Time 2 into the variables windows and press OK). We would get the output shown in Table 25.5). Here we can see that our academic vindictiveness scale shows acceptable test–retest reliability with a correlation of $r = 0.77$, larger than the criteria of $r = 0.7$.

Validity

Validity is concerned with whether a test is measuring what we claim it is measuring. Therefore if we're proposing that the 18 items developed above measure academic vindictiveness, how can we show this is a measure of academic vindictiveness? There is no absolute way of showing that it is a measure of academic vindictiveness; rather, what we would have to do is try to collect evidence through a number of criteria to support the validity of our test as a measure of academic vindictiveness. Traditionally, a number of validity criteria can be applied to psychometric tests.

Stop and think

? ! ?

The difference between using a psychometric tests as a diagnostic test and as a variable

Psychometric tests are used as a powerful assessment of clinical states. Here, we're going to give an outline of how these are and should be used. One example of a clinical assessment instrument is the Edinburgh Postnatal Depression Scale. The ten-item scale was developed by Cox, Holden and Sagovsky (1987) as a way of assisting primary care health professionals to detect Postnatal Depression, thought to affect at least 10 per cent of women, among mothers who have recently had a baby. Before using the scale, researchers need to be aware of the following three things:

1 Care should be taken to avoid the possibility of the mother discussing her answers with others.
2 The mother should complete the scale herself, unless she has limited English or has difficulty with reading.
3 The EPDS may be used at 6–8 weeks to screen postnatal women. The child health clinic, postnatal check-up, or a home visit may provide suitable opportunities for its completion.

The scale has ten questions. But before going on to the questions, any respondent is given the following the instructions:

1 The mother is asked to underline the response which comes closest to how she has been feeling in the previous seven days.
2 All ten items must be completed.

Respondents will be then asked to complete the following scale.

As you have recently had a baby, we would like to know how you are feeling. Please UNDERLINE the answer which comes closest to how you have felt IN THE PAST 7 DAYS, not just how you feel today.

1 I have been able to laugh and see the funny side of things.

As much as I always could
Not quite so much now
Definitely not so much now
Not at all

2 I have looked forward with enjoyment to things.

As much as I ever did
Rather less than I used to
Definitely less than I used to
Hardly at all

3 *I have blamed myself unnecessarily when things went wrong.

Yes, most of the time
Yes, some of the time
Not very often
No, never

4 I have been anxious or worried for no good reason.

No, not at all
Hardly ever
Yes, sometimes
Yes, very often

5 *I have felt scared or panicky for no very good reason.

Yes, quite a lot
Yes, sometimes
No, not much
No, not at all

6 *Things have been getting on top of me.

Yes, most of the time I haven't been able to cope at all
Yes, sometimes I haven't been coping as well as usual
No, most of the time I have coped quite well
No, I have been coping as well as ever

7 *I have been so unhappy that I have had difficulty sleeping.

Yes, most of the time
Yes, sometimes
Not very often
No, not at all

8 *I have felt sad or miserable.

Yes, most of the time
Yes, quite often
Not very often
No, not at all

9 *I have been so unhappy that I have been crying.

Yes, most of the time
Yes, quite often
Only occasionally
No, never

10 *The thought of harming myself has occurred to me.

Yes, quite often
Sometimes
Hardly ever
Never

Answers are scored 0, 1, 2, and 3 according to increased severity of the symptoms. So for Item 1 'I have been able to laugh and see the funny side of things'. 'As much as I always could' would be scored 0 and 'Not at all' would be scored 3, meaning a higher score '3' would mean the respondent had not at all been able to laugh and see the funny side of things indicating they may be depressed. Items marked with an asterisk are reverse scored (though they would not be marked on the scale that is administered to participants). So, for example, for the item 10, 'The thought of harming myself has occurred to me', 'Yes quite often' would be scored as 3 and 'Never' as 0. A total score for the scale is then computed by adding together the scores.

In research terms, researchers would use an overall score on the scale for use in correlational analysis. However, these scales can also be used to make an assessment. In terms of possible score, they could range from a minimum of 0 (not depressed at all) to 30 (very depressed) across the 10 items. It is unlikely that many people will score 30 on the scale. Indeed Cox *et al*. (1987) suggest that anyone scoring above 10 or greater may be suffering from possible depression. They also suggest that any assessment should also look at item 10 closely for any possible suicidal thoughts.

This is the main thing about these sorts of questionnaires; they are a tool to make an initial assessment. Cox *et al*. emphasise that the scale should not be used over clinical judgement and is only a tool, and where mothers score above 10 then a deeper clinical assessment should be performed to confirm the diagnosis. This is true of many of these types of tests; they are best used as potential indicators, but should never be used as a single indicator of anything.

There is one important distinction to make when using the test. Often, when using a psychometric test to measure any attitude or behaviour there is a temptation to try to categorise people as falling into a group having the attitude behaviour, or falling into a group who do not demonstrate that behaviour. However, rarely in research do researchers make this distinction. They will most often just use scores on the test as a variable, and try not to categorise people (unless they have a very good reason to do so). This is because, as we have highlighted above, it is extremely difficult to make a diagnosis (i.e. group someone) just using a test alone. Indeed for many tests it would be impossible to categorise someone for a behaviour. What is the criteria threshold for deciding if someone is intelligent or not, neurotic or not, or optimistic or not. Consequently in individual difference and intelligence research we tend to look at scores on scales, rather than splitting our samples into diagnostic groups, realising that clinical expertise is needed before making such distinctions.

These different types of validity include:

- **Convergent validity** – A psychometric test's convergent validity is assessed by the extent to which it shows associations with measures that it should be related to. So, for example, our new academic vindictiveness should be related to other aspects of vindictiveness; for example, a tendency to seek revenge, show spitefulness and vengefulness, particularly in academic settings.
- **Concurrent validity** – A psychometric test is thought to show concurrent validity when it shows acceptable correlations with known and accepted standard measures of that construct. Therefore, it is slightly different to convergent validity because it isn't against other related criteria, but criteria that are reportedly measuring the same thing. So, we would expect our new academic vindictiveness to be related to other available measures of academic vindictiveness that already exist. However, sometimes this is difficult to assess if there are no other measures of the construct.
- **Discriminant validity** – Something shows discriminant validity when it is *not* related to things that it shouldn't be related to. Sometimes this is difficult to assess because the finding needs to be useful. For example, there is little point suggesting that academic vindictiveness should not be related to cake-eating, because that tells us very little about the construct. Therefore, when sometimes examining the discriminate validity of a construct, researchers suggest that the new construct should *not* share high correlations with other constructs. For example, a measure of academic vindictiveness should not share a high correlation with any of the main five personality dimensions (neuroticism, agreeableness, extraversion, openness and conscientiousness). However, we might expect that the academic vindictiveness scale to share a small negative correlation with agreeableness, suggesting that the scale is largely independent of the five-factor model of personality).
- **Face validity** – This aspect of validity is concerned with what the measure *appears* to measure. Therefore, when a test has face validity, it means that it does look like a test that measures the concept it is designed to measure. As in the case of our new academic vindictiveness, the best way to consider this question would be to group together

some experts in academic vindictiveness to judge whether they think the questionnaire represents a good measure of that construct.

- **Content validity** – refers to the extent to which a measure represents all facets of the phenomena being measured. So, for example, in the case of academic vindictiveness we might have some ideas about different types of academic vindictiveness. For example, there might be academic vindictive behaviours, academic vindictive attitudes and academic vindictive feelings. Therefore, does our measure of academic vindictiveness cover all these domains, i.e. does the content of the test try to represent all aspects of academic vindictiveness (i.e. behaviours, attitudes and feelings)? If it doesn't, then it cannot be said to have content validity.
- **Predictive validity** – This type of validity assesses whether a measure can accurately predict something in the future. For example, in the case of academic vindictiveness, it should be able to predict people acting in an academically vindictive way in the future. Therefore, our researcher might administer our new academic vindictiveness to a group of students at the beginning of the academic year and then measure the extent to which individuals reported a number of academic vindictive behaviours during the examination period at the end of the semester (for example, not sharing notes, not helping other people revise). If our new academic vindictiveness demonstrated predictive validity, it would be able to predict those students who didn't share notes or help other people revise during the examination period.
- **Third person rating of the individual** – Getting other people to rate the individual on the items of the questionnaire is a very good way of potentially assessing the validity of the questionnaire because, ideally, the ratings given by the participants and the other person should be similar. Indeed sometimes it is referred to as the gold standard of validity testing. Here, you might ask a person close to the individual filling in the questionnaire (e.g. a friend or member of family) to rate that individual on each of the items of the questionnaire. Therefore if the person's ratings of the individual for each questionare similar to the person's actual answers, then this provides an independent assessment of the questionnaire and validity for the questionnaire. So, for example, for our academically vindictive questionnaire, we would expect scores for individuals to be correlated with their friends and family ratings of them.

However, psychometricians have developed some of these ideas into wider terms, and commonly in the modern psychometric literature you will see reference to two further ideas: *Criterion-related validity* and *Construct validity*. Both these forms of validity combine different types of validity we have just mentioned to form new concepts of validity.

The first, criterion-related validity, assesses the value of the test by the respondents responses on other measures, i.e. criteria. Two types of measure fall under criterion-related validity, **concurrent validity** (assesses whether the measure shows acceptable correlations with known and accepted standard measures of that construct) and **predictive validity** (assesses whether a measure can accurately predict something in the future).

The second is **construct validity**. Construct validity as a more general term and refers to validity that seeks to establish a clear relationship between the construct at a theoretical level and the measure that has been developed. Construct validity can be informed by different types of validity previously mentioned, but in 1959 two authors, Donald Campbell and Donald Fiske (Campbell and Fiske, 1959), introduced that two types of validity: convergent validity (that the measure shows associations with measures that it should be related to) and discriminant validity (that the measure is *not* related to things that it should not be related to) as the subcategories of construct validity and they developed a theory named the multitrait-multimethod which seeks to assess the construct validity of a measure by balancing the assessments between convergent validity and discriminant validity. You can read more about this method in the Stop and think box: Multitrait-multimethod matrix (page **000**).

So let us test the validity of the academic vindictiveness scale using the following validity criteria. In this section we're going to present a number of validity checks to see to what extent our new scale shows validity as a good measure of academic vindictiveness. What is important to note is that there is currently no other measure of academic vindictiveness so we cannot carry out a concurrent validity check. We therefore need to look at the other forms of validity.

The first is convergent validity, which assesses the extent to which our scale shows associations with measures that it should be related to. In this case, we could look at some other measure of traits that a person who is academically vindictive should also score high on. For example, the International Personality Item Pool (see Stop and think: International Personality Item Pool) provides public domain (free to use) access to a number of personality and individual difference measures. For example, the site provides trait measures of *Morality* (e.g. 'I would never scheme against others'), ***Integrity/Honesty*** (e.g. 'I can be trusted to keep my promises') and *Machiavellism* (a trait that suggests that deceit is justified in pursuing and maintaining power, e.g. 'Find it easy to manipulate others'). Therefore, we would expect a significant positive correlation between academic vindictiveness and Machiavellism and a significant negative correlation between academic vindictiveness and morality and integrity/honesty.

Stop and think

? ! ?

International Personality Item Pool

The International Personality Item Pool (IPIP) website (**http://ipip.ori.org/ipip/**) provides public domain access to a number of personality and individual difference measures. The main reason for the development of the site is that a lot of measures of personality (the MMPI, 16PF and NEO-PI) are copyrighted by the test authors and consequently cannot be used freely by other researchers, like you. The site provides freely available versions of all the main personality measures as well as over 200 measures of personality individual difference traits.

In terms of discriminant validity a good strategy would be to compare the academic vindictiveness scale to the five-factor personality measures; neuroticism, extraversion, agreeableness, conscientiousness and openness. Now, while we might expect academic vindictiveness to share a negative correlation with agreeableness (a tendency to be pleasant and accommodating), we can show discriminate validity for our measure by finding no significant relationship between our measure of academic vindictiveness and neuroticism, extraversion, conscientiousness and openness.

Our final attempt at validity is to examine third person ratings of the individual on the scale. Here, we asked a friend of the participant to rate the participant on the measure of academic vindictiveness. However, for this measure we asked the respondent to provide a friend who was also a classmate, so as to ensure that the friend had some knowledge of how the respondent was in an academic setting.

We now have a number of measures to establish validity of our scale. All these measures are included in the data set we used earlier, and if you wish you can run your own correlational analysis (remember we showed you how to run a correlation in the test – retest example, and remember our academic vindictiveness scale variable name in the data set is ACADEMVIND). However, Table 25.6 shows the correlations between our academic vindictiveness scale and of morality, integrity/honesty and Machiavellism, the five-factor model of personality and the friend's rating of the respondent on the scale.

Our first set of results tries to establish convergent validity for our academic vindictiveness scale. In terms of our predictions, academic vindictiveness shares a significant negative correlation with morality and integrity/honesty and a significant positive correlation with Machiavellism (see Table 25.6). What is also worth noting here is the size of the correlations. You remember that the effect size of a correlation statistic of $r = 0.1$ is viewed as small, $r = 0.3$ as medium (or moderate) and $r = 0.5$ as large. Here, the correlations are at least of a medium size.

We also sought to examine both the convergent but mainly the discriminate validity of our academic vindictiveness scale by looking at the relationship between the scale and the five-factor model of personality (see Table 25.7). Again, the findings are consistent with our predictions. Academic vindictiveness shares a significant negative correlation with agreeableness (with a medium effect size), but our discriminate validity is demonstrated by its lack of any significant correlation with neuroticism, extraversion, openness and conscientiousness.

Table 25.6 Pearson product moment correlation coefficients between academic vindictiveness and morality, integrity/honesty and machiavellism.

	Academic vindictiveness
Morality (IPIP)	−0.453**
Integrity/honesty (IPIP)	−0.410**
Machiavellism (IPIP)	0.344**

**$p < 0.01$

Table 25.7 Pearson product moment correlation coefficients between academic vindictiveness and the five-factor personality dimensions.

Neuroticism	0.076
Extraversion	0.013
Openness	−0.074
Agreeableness	−0.320(**)
Conscientiousness	−0.097

**$p < 0.01$

Table 25.8 Pearson product moment correlation coefficients between academic vindictiveness and peer rating.

Peer rating	0.501**

**$p < 0.01$

Finally, we had asked a friend to rate the person on the measure of academic vindictiveness (see Table 25.8). As we can see, there was a positive significant correlation between the respondent's answers and the friends rating, with a large effect size.

Together we can see that there is some validity for our measure, particularly by the scales' expected relationships to measures of morality, integrity/honesty, agreeableness and Machiavellism as well as a clear association with a classmate and friend's rating.

This is the main point of establishing the validity of a test. You can never ascertain that it is perfectly valid, but you can carry out different studies and tests to develop more and more evidence to support the validity of your test. If you can replicate findings and find further ways of testing the validity of a scale you can build up the evidence base for assessing the validity of your test.

Stop and think

? ! ?

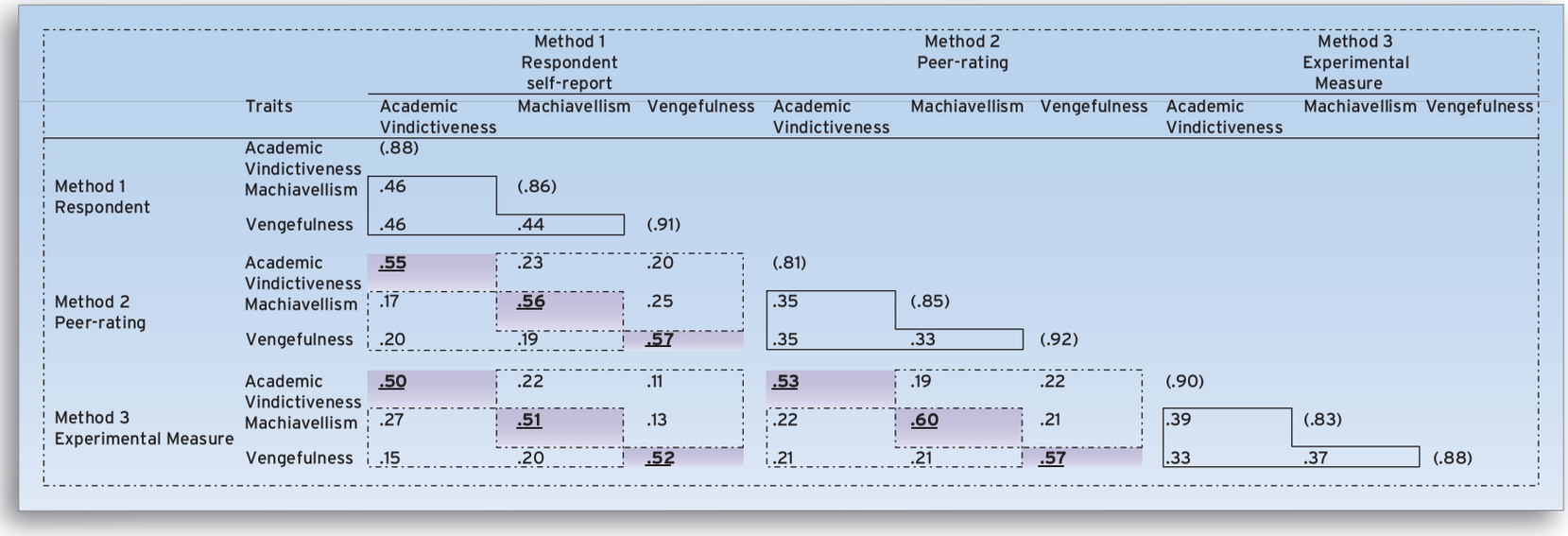
Multitrait-multimethod matrix

The multitrait-multimethod matrix establishes the construct validity of two or more constructs by two or more methods of assessment. What the multitrait-multimethod matrix does is establish the convergent and discriminant validity for a measure by comparing the level of correlations between the different constructs by the different methods. Usually this information is presented alongside reliability statistics to provide a rich assessment of the reliability and overall construct validity for a measure. Therefore we could demonstrate the multitrait-multimethod matrix with three different methods of measuring three different constructs in relationship to our academic vindictiveness scale. For our current example, let us suggest that in addition to academic vindictiveness we could measure two potentially positively related constructs to academic vindictiveness: Machiavellism and vengefulness. We would then seek to measure each of these constructs by three methods:

- Self-report psychometric test (Method 1).
- Peer-ratings of each of these constructs by a friend (Method 2).
- Behaviour measures in relation to an experimental task (Method 3).

We have outlined a traditional way in which the results of these measures would have then been examined using the multitrait-multimethod matrix. This matrix is presented in Figure 25.3 and provides us with information on convergent validity and discrimnant validity.

The first thing to note in this analysis is the figures highlighted in parentheses. These would normally be the reliability statistics for each of the tests. This could be any indicator or reliability (Cronbach's or test–retest). However, we are less concerned with these statistics at this stage; we would hope to see high reliability statistics in this presentation. The rest of the statistics presented in the matrix are correlational statistics. The next thing to note is those figures surrounded by a black line. These are the correlations between the different constructs as measured by the same method. When we are looking at similar constructs we would expect relatively a positive correlation between these constructs, particularly for our academic vindictiveness measures (as we hypothesised these constructs should show convergent validity). However, it is the other sets of figures we are most concerned with in terms of the multitrait-multimethod matrix and considering convergent validity and discrimnant validity. Those correlations with a grey background and bold and underlined figures are the same construct as measured by three different measures. These correlations should be positive and strong (i.e. relatively large) and provide us with our evidence of convergent validity (i.e. the self-report measure of academic vindictiveness should share the strongest correlations with our peer-rating and experimental measures of academic vindictiveness). The constructs surrounded by the dotted line are different constructs as measured by different measures, and it is these correlations that should be the lowest of all within the matrix as they supply us with evidence of discriminant validity (i.e. an experimental measure of vengefulness should share one of the lowest correlations with the self-report measure of academic vindictiveness). Together these findings of convergent validity and discrimnant validity provide one of the most eloquent and comprehensive ways of providing construct validity (convergent and discrimnant validity) for a psychological test.



	Traits	Method 1 Respondent self-report			Method 2 Peer-rating			Method 3 Experimental Measure		
		Academic Vindictiveness	Machiavellism	Vengefulness	Academic Vindictiveness	Machiavellism	Vengefulness	Academic Vindictiveness	Machiavellism	Vengefulness
Method 1 Respondent	Academic Vindictiveness	(.88)								
	Machiavellism	.46	(.86)							
	Vengefulness	.46	.44	(.91)						
Method 2 Peer-rating	Academic Vindictiveness	**.55**	.23	.20	(.81)					
	Machiavellism	.17	**.56**	.25	.35	(.85)				
	Vengefulness	.20	.19	**.57**	.35	.33	(.92)			
Method 3 Experimental Measure	Academic Vindictiveness	**.50**	.22	.11	**.53**	.19	.22	(.90)		
	Machiavellism	.27	**.51**	.13	.22	**.60**	.21	.39	(.83)	
	Vengefulness	.15	.20	**.52**	.21	.21	**.57**	.33	.37	(.88)

Figure 25.3 A multitrait-multimethod matrix of a self-report measure of academic vindictiveness, peer rating of Machiavellism and an experimental measure of Vengefulness.

Profile

Paul Kline

Professor Paul Kline was born in 1937. He became an academic psychologist after training as an educational psychologist, which followed a period in which he taught classics before moving on to completing his PhD at Manchester University. In 1969, he undertook a position in the Exeter University Psychology department, and he later became the first professor of Psychometrics in the United Kingdom.

It's been noted by his colleagues that Kline, whose untimely death was in 1999, had two enthusiasms in psychology – psychometrics and Freudian theory. He wrote essential books on psychometrics. His 1986 handbook of test construction provided an essential and clear introduction to a complex field. His 1999 book was essential in that it covered psychometric theory, the different kinds of psychological tests and applied psychological testing as well as evaluating the best-published psychological tests. His 2000 book, *The New Psychometrics: Science, Psychology and Measurement*, sought to use his knowledge of psychological measurement to argue that truly scientific forms of measurement could be developed to create a new psychometrics that would transform psychology from a social science to a pure science. However, perhaps, in his most famous book, *Fact and Fantasy in Freudian Theory*, he was able to combine statistics and Freudian theory and brought forward many principles of reliability and validity to examine psychological studies of Freudian theory.

Stop and think

Self-assessment exercise

In our validity assessment of the academic vindictiveness scale we didn't examine the face validity and the predictive validity of our test. Can you think of an empirical study that would test the academic vindictiveness scale for both these sets of validity?

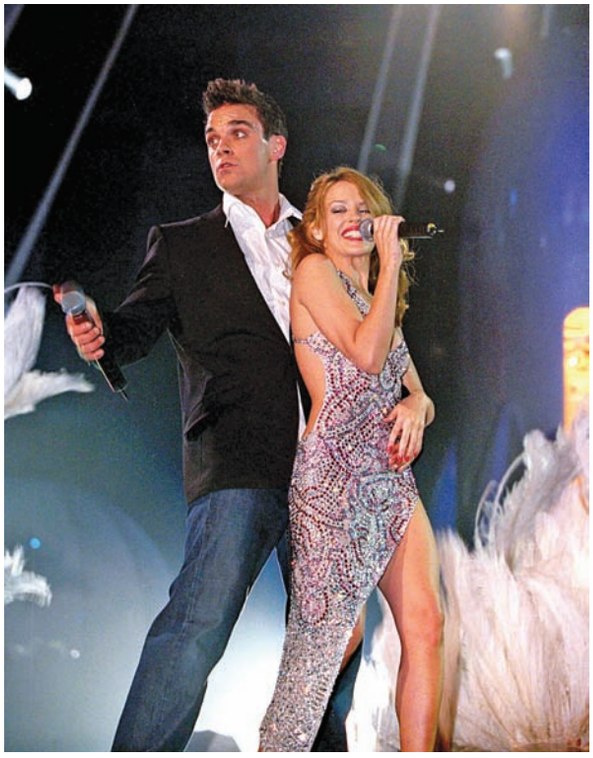
Advanced techniques in psychometric evaluation: factor analysis

In this section we are going to introduce you to an advanced analysis technique that can be used to look at the psychometric properties of items. Factor analysis, in a psychometric test context, refers, in the first instance, to the area that we covered within the internal reliability of a test earlier in this chapter. In this part of this chapter we are very interested in seeing whether a number of items correlated together to form a single scale. However, what happens if the scale you've developed is designed to have more than one element to it? For example, personality measures have a number of factors; the five-factor personality scale has five factors. What happens if a number of elements exist to our academic vindictiveness measure, not just one? Well, one technique that is used to explore these issues is factor analysis. In this section we're going to explain two types of factor analysis to you: exploratory factor analysis and confirmatory factor analysis. Exploratory factor analysis (EFA) is used for exploring what elements (or as they are known, factors) underlie sets of items. Confirmatory factor analysis (CFA) is used to try to confirm what you may have found with exploratory factor analysis.

Factor analysis

Factor analysis is a multivariate (multiple variables) 'data reduction' statistical technique that allows us to simplify the correlational relationships between a number of variables.

Why would you wish to do so? Imagine if you wanted to look at the relationships between variables 1 to 20. This would imply having to interpret 190 relationships between variables. (For example, variable 1 can be correlated with variables 2 through 20. And, variable 2 can be separately associated with variables 3 through 20, variable 3 with variables 4 through 20 and so on.) Identifying real patterns is complicated further because the relationship between, for example, variables 1 and 12 may be affected by the separate relationships each of these variables has with variable 13, with variable 14 and so on. This complicated explanation of 190 relationships is a nightmare for researchers. The researcher will find it difficult to explain which variable is actually related to which other variables, as they may be uncertain whether the apparent relationship between two



Sometimes things that are separate overlap.
Source: Alamy Images

variables is genuine, or simply a facet of both variables' relationships with another third variable. Indeed, even writing an explanation of multiple correlations is difficult. What factor analysis does is provide reliable means of simplifying the relationships and identifying within them what *factors*, or common patterns of association between groups of variables, underlie the relationships.

Factor-analytic techniques are a solution to this type of problem, allowing us to look for simple patterns that underlie the correlations between many variables. Yet you may be surprised that you already use factor analysis regularly in your life. Imagine all musicians and music groups in the world. Now, think about the different definitions you can apply to sets of these groups. Some of these artists are very similar; some are very different. However, with the vast majority, you can categorise them as either Pop Music Artists, Rap Artists or Jazz, Dance, R&B and so on. What you're doing through this categorising is simplifying a wealth of information regarding music to aid your understanding. So, when someone asks you what sort of music you like, rather than listing lots of groups – Destiny's Child, Justin Timberlake, Jennifer Lopez, Kanye West, Mary J. Blige, Kelis – you might just simply say, 'R&B'. Well, factor analysis is very similar to this.

Let us take you through an example of factor analysis. Imagine the following 10 musical artists: Royksopp, BassHunter, Prodigy, 50 Cent, Snoop Dogg, Franz Ferdinand, Coldplay, Keane, Miles Davis and Charlie Parker. If we were to perform a factor analysis on these music artists, and the extent to which people like them, we might see something like Table 25.9.

Table 25.9 Imaginary factor analysis of musical artists.

	Factor 1	Factor 2	Factor 3	Factor 4
Royksopp	**0.73**	−0.21	0.09	0.21
BassHunter	**0.67**	−0.11	0.12	0.07
Prodigy	**0.55**	−0.05	0.03	0.02
50 Cent	0.10	**0.81**	0.18	0.03
Snoop Dogg	0.02	**0.85**	0.21	0.05
Franz Ferdinand	0.03	0.23	**0.74**	−0.01
Coldplay	0.14	0.06	**0.65**	0.02
Keane	0.12	0.10	**0.55**	−0.04
Miles Davis	0.25	0.05	0.01	**0.56**
Charlie Parker	0.20	0.02	−0.03	**0.78**

What the factor analysis does is determine, first, the number of factors that exist. As you can see, our imaginary factor analysis has four factors. Then, what factor analysis does is determine where on which factor each variable (i.e., each artist) falls by a number. This number, called a loading, can be positive or negative and can range between −1.00 through to +1.00. Regardless of whether the number is positive or negative, the higher the number is on the factor, the more important the variable is to the factor. Kline (1986) has suggested that you should ignore any number less than 0.3, and other authors have suggested that those numbers of above 0.4 are important.

Using the imaginary factor analysis of music artists, we have shown the loadings above 0.4 in bold. From this analysis we would argue that the music artists broke down into four factors, the first factor being Dance (Royksopp, BassHunter, Prodigy), the second being Rap (50 Cent, Snoop Dogg), the third being Indie (Franz Ferdinand, Coldplay, Keane) and the fourth being Jazz (Miles Davis and Charlie Parker).

You will find reference to factor analysis throughout this book. However, there are two main ways you will see a reference to factor analysis: factor analysis that deals with (1) simplifying relationships between single questions or items and (2) simplifying relationships between variables). For (2), simplifying relationships between variables, turn to Chapter 12 on the intelligence and the theory of 'g' to see how factor analysis has been applied. But in this chapter we are going to use exploratory factor analysis for (1) simplifying relationships between single questions or items. You will commonly see factor analysis used on items on a scale. All scales will have several items or questions, and factor analysis can be used to understand the relationships between these items. Therefore, you will see a factor analysis when researchers have devised a new psychological measure, or want to examine existing psychological measures

and want to see how the items on that scale break down into different factors, or indeed whether they form one factor. In the next section we are going to use a research example to show you how to carry out an exploratory factor analysis on a set of items to develop a psychometric scale.

Exploratory factor analysis

To illustrate exploratory factor analysis we are going to use the example of our academic vindictiveness scale and examine the factor structure of the scale. There are two main steps to exploratory factor analysis which we're going to outline here.

1 To determine how many factors underlie our data. This procedure is called extraction of factors.
2 To determine which items/variables load on each of the factor. This procedure is called rotation of factors.

Extraction

Extraction techniques allow you to determine the number of factors underlying the relationship between a number of variables. There are many extraction procedures, but the most common techniques, and the one we are going to use is actually called principal component analysis ('factors', when using principal components analysis, are actually called components, but for simplicity we will continue to call them factors).[2]

Now, in terms of determining the number of factors there are three methods. The first two methods we can obtain from SPSS for Windows. Load up the data set you used previously in this chapter. Pull down the **Statistics** menu, by clicking **Analyze** and then click on **Data Reduction** and then **Factor**. You should then get a screen that looks like Figure 25.4.

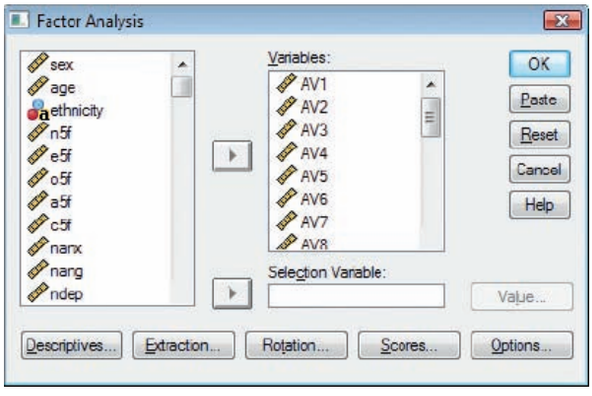


Figure 25.4 Factor analysis window.

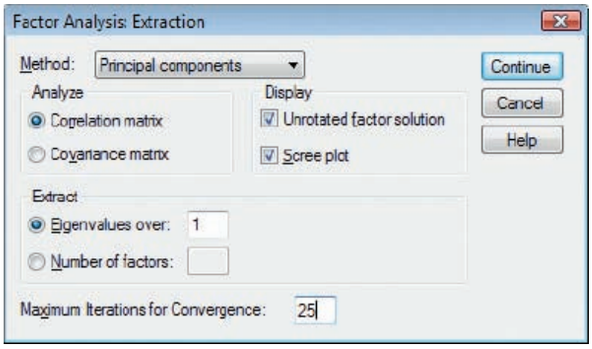


Figure 25.5 Factor analysis: extraction window.

Transfer your variables into the **Variables** box (remember to enter only variables AV1 to AV18) and then click on **Extraction**. You should then get a screen that looks like Figure 25.5.

Click on the box next to **Scree plot** and then press **Continue** and then **OK**. You should get an output that contains the table and figure shown in Table 25.10.

The first technique for deciding on the number of factors can be gained from the table. Here we are interested in the second column 'Total' under the initial eigenvalues. We can see a list here, 6.713, 1.538, 0.861, 0.841 etc. Here the number of factors extracted are the number of values above one, and in our case two eigenvalues, 6.713, 1.538 are above one. This is a traditional way of deciding factors. SPSS calculates the number of factors, and assigns in descending order an eigenvalue. Traditionally, factors with eigenvalues above 1 are seen as significant factors and SPSS will extract that number of factors. However, some statisticians claim that using eigenvalues can be unreliable and use something called the scree test. What the Scree test does is plot the eigenvalues so that use a visual assessment to see how factors should be extracted (see the Figure 25.6).

What the Scree plot does is plot the eigenvalues and we are meant to use a visual criterion to determine the number of factors. The Scree plot is named after the debris that collects at the bottom of a rocky slope on a mountain (a scree). This is important as we can see the Scree plot looks like the side of a mountain. We determine the number of factors by selecting those eigenvalues that occur before the plot straightens outs (or rough straightens out). Another way is to imagine the plot as an arm with a bend in the elbow. You would select all points above the elbow. In this case the Scree plot flattens out (or the elbow occurs) at the third point, so we take all points before that, i.e. 2 points. This is the number of factors we should extract.

[2]The difference between principal component analysis and factor analysis (though the procedures are the same) is that principal components analysis is used to simplify correlations between variables, while factor analysis is concerned with underlying factors to the correlations between variables.

Table 25.10 Extraction output.

Total variance explained

	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings*
Component	Total	% of Variance	Cumulative %	Total	% of variance	Cumulative %	Total
1	6.713	37.295	37.295	6.713	37.295	37.295	5.805
2	1.538	8.542	45.837	1.538	8.542	45.837	5.468
3	0.861	4.784	50.621				
4	0.814	4.520	55.141				
5	0.775	4.303	59.444				
6	0.756	4.200	63.644				
7	0.729	4.049	67.693				
8	0.697	3.870	71.563				
9	0.666	3.702	75.265				
10	0.621	3.449	78.713				
11	0.599	3.328	82.042				
12	0.545	3.029	85.071				
13	0.531	2.949	88.020				
14	0.505	2.807	90.827				
15	0.471	2.618	93.445				
16	0.412	2.288	95.733				
17	0.393	2.182	97.915				
18	0.375	2.085	100.000				

Extraction method: principal component analysis.
*When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

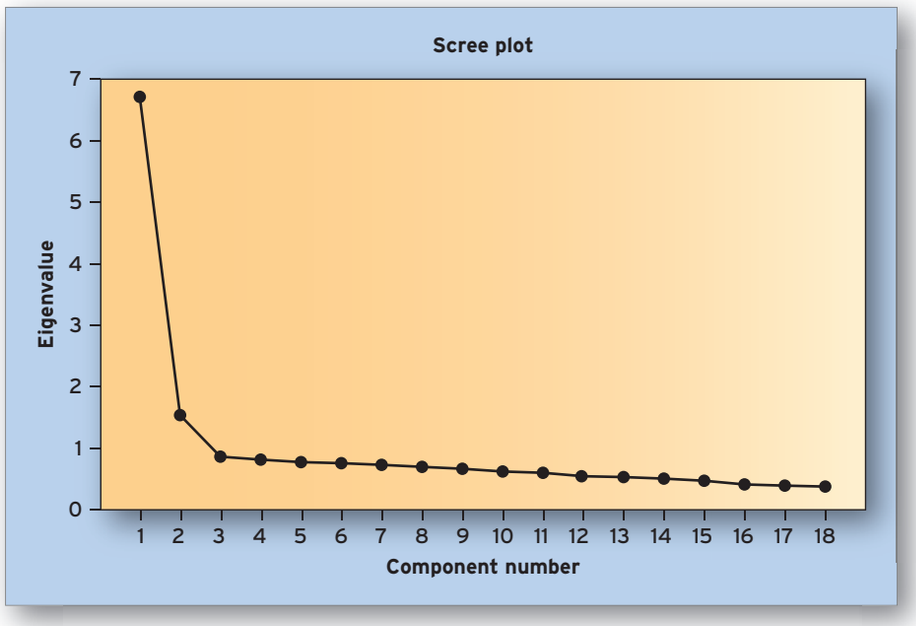


Figure 25.6 Scree plot.

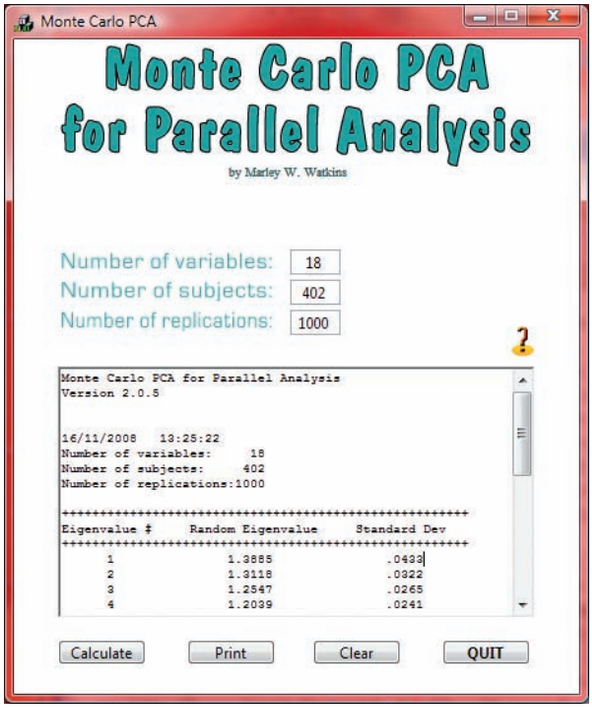


Figure 25.7 Monte Carlo PCA for parallel analysis.

So far both techniques have suggested two factors be extracted. However, a third assessment, and one that is often seen as most accurate, is something called Parallel Analysis of Monte Carlo simulations (Horn, 1965). What this technique does is create a data set of random sets of eigenvalues that would be expected from purely random data with no structure. In simple terms, in this analysis you are comparing your data set against a data set that would happen completely by chance and therefore this analysis tells you what is important and what are just chance findings. In parallel analysis you compare the eigenvalues of your data set against the eigenvalues of the random data set to determine the number of factors to be extracted.

Now, the software for creating this data set is not in SPSS for Windows, but is a small simple program that is free and downloadable.[3] If you install it on your machine and run it you'll get a simple interface, as in Figure 25.7 (though if you are running this in a university computer lab the computer might not let you run the software).

The program is simple, and all you need to input is **Number of Variables**, **Number of Subjects** and **Number of Replications**. For our data we have 18 variables, 402 respondents so we input this into these boxes (see Table 25.11). We also enter 1000 into the number of replications, this is just a standard figure (however, if your computer has problems calculating this and freezes lessen the number in this box to 100).

In the empty box of the application you should get the following (on the top of Figure 25.7). This is the criteria by which we will judge our own eigenvalues, these are provided at the bottom of Table 25.11.

Table 25.11 Monte Carlo Analysis and the data set eigenvalues.

Monte Carlo PCA for Parallel Analysis
Version 2.0.5

16/11/2008 13:25:22
Number of variables: 18
Number of subjects: 402
Number of replications: 1000

Eigenvalue #	Random eigenvalue	Standard dev
1	1.3885	0.0433
2	1.3118	0.0322
3	1.2547	0.0265
4	1.2039	0.0241
5	1.1597	0.0223
6	1.1183	0.0197
7	1.0797	0.0191
8	1.0418	0.0183
9	1.0051	0.0182
10	0.9694	0.0178
11	0.9346	0.0174
12	0.9001	0.0182
13	0.8653	0.0185
14	0.8309	0.0196
15	0.7951	0.0196
16	0.7573	0.0207
17	0.7167	0.0222
18	0.6671	0.0275

Monte Carlo PCA for Parallel Analysis

Our eigenvalues . . .

Total variance explained

	Initial eigenvalues		
Component	Total	% of Variance	Cumulative %
1	6.713	37.295	37.295
2	1.538	8.542	45.837
3	0.861	4.784	50.621
4	0.814	4.520	55.141
5	0.775	4.303	59.444

Extraction method: principal component analysis.
When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

[3]The program is available at **http://www.softpedia.com/get/Others/Home-Education/Monte-Carlo-PCA-for-Parallel-Analysis.shtml** or type 'Monte Carlo PCA for Parallel Analysis' into Google.

We then compare each of our eigenvalues against the random eigenvalues in turn. So we compare the first of our eigenvalues against the first random eigenvalues, and then each one after that. At first, our eigenvalues exceed those in the random data set, but when they stop exceeding the ones in the data set that's our criteria for determining the number of factors. For example, our first eigenvalue '6.713' exceeds the first eigenvalue in the random data set '1.3885'. Similarly, our second eigenvalue '1.538' exceeds the second eigenvalue in the random data set, '1.3118'. However, our third eigenvalue, '0.861' fails to exceed the third eigenvalue in the random dataset, '1.2547'. This is our criteria for determining those factors, we only extract those factors that are greater in value than corresponding values in the data set. That is, because the generated data set is of purely random data, we are only sure that those eigenvalues that exceed the value in the random data set are not created by chance. Here two of our eigenvalues exceed the values in the random data set, and therefore we would select two factors.

Among our data set all three criteria suggest that two factors be extracted from our data set.

Rotation

Rotation is necessary when extraction techniques suggest there are two or more factors to extract. Simply put, the rotation of factors is designed to give us an idea of how the factors we extract are related, and provides a clear picture of which items load on which factor.

There are two sets of rotations techniques.

1. **Orthogonal rotations** – These are rotations that assume that each factor shares no association and are specifically unique. This is often used in psychology when applying a theoretical model to factor analysis and the model predicts that the factors are independent.
2. **Oblique Rotations** – These tend to be used more often, as this procedure outlines the position of factors to one another, i.e. determines how they are usually related. The most specific procedure most often used for this is a technique called oblimin.

Each of these categories has a number of different types of rotation within them. However, the most specific procedure most often used for orthogonal rotations is a technique called varimax. The most specific procedure most often used for oblique rotations is a technique called oblimin. Usually, there are reasons for doing one or another. For example, researchers tend to prefer oblique rotations because they actually indicate how the different factors are related to one another. However, researchers will use orthogonal rotations when they feel the factors should be independent. Because of these differing views, and this would make sense when exploring factors, researchers tend to use both and report on those that produce the clearest results. We will now take you through an example that uses both these rotation techniques to illustrate rotation to you.

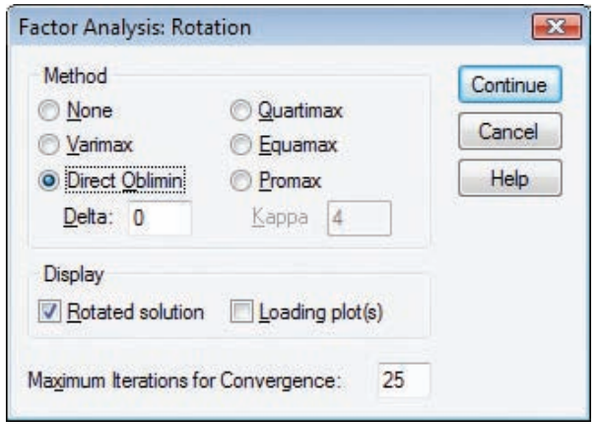


Figure 25.8 Factor analysis: rotation window.

OK, we need to run our factor analysis procedure again as before, but this time in the Factor Analysis Window we press the **Rotation** button. You will then get a Window that looks like Figure 25.8. Click on **Direct Oblimin** and press **Continue** and then **OK**. Then run the whole procedure again, but this time click on **Varimax**. Then run the whole procedure again.

You should have two outputs that contain the tables shown in Figure 25.8. For the oblimin rotation the output you should look at is the pattern matrix (Table 25.12a). For the varimax rotation the output you should look at is the rotated component matrix (Table 25.12b). We have included the items so you can accurately see the item. In the SPSS output you will just have the variable names.

First, we're going to look at the first table (Table 25.12a), the pattern matrix for the oblimin rotation. Each item loads on a factor. Loadings are the strength of the variable in defining the factor. Loadings on factors can be positive or negative, and when a number happen together this means very little to us at this stage. For instance, in the present example, the large majority of the loading on the first factor are positive and the large majority of the loadings on the second factor are negative. However, an item with a negative loading on a factor with a lot of items with a positive loading (or vice versa) would indicate that this variable has an inverse relationship with these other variables on the factor.

Now what we have to determine is what factor which item loads on. Opinions are fairly arbitrary about the point at which an item loads on a factor. However, Child (2006) suggests that anything above 0.44 could be thought of as important, with increased value of the loading becoming more vital in determining what the factor is Kline (1986) suggests that you should not ignore any loading below 0.3. More commonly a high loading (i.e. 0.40 or higher) on one

Table 25.12 Pattern matrix (a) and rotated component matrix (b).

(a) Pattern matrix

	Component	
	1	2
AV1	−0.072	−0.745
AV2	−0.029	−0.799
AV3	−0.044	−0.753
AV4	−0.003	−0.677
AV5	0.645	−0.125
AV6	0.097	−0.601
AV7	0.017	−0.558
AV8	0.546	−0.146
AV9	0.542	−0.053
AV10	0.684	0.121
AV11	0.764	0.031
AV12	0.630	−0.124
AV13	0.672	0.010
AV14	0.183	−0.480
AV15	0.595	−0.048
AV16	0.081	−0.722
AV17	0.749	0.074
AV18	0.597	−0.031

Extraction method: principal component analysis.
Rotation method: oblimin with Kaiser normalisation.
Rotation converged in five iterations.

(b) Rotated component matrix

	Component	
	1	2
AV1	0.159	0.687
AV2	0.216	0.752
AV3	0.187	0.704
AV4	0.203	0.644
AV5	0.651	0.319
AV6	0.275	0.603
AV7	0.186	0.536
AV8	0.564	0.309
AV9	0.531	0.219
AV10	0.613	0.097
AV11	0.716	0.208
AV12	0.636	0.314
AV13	0.636	0.200
AV14	0.320	0.514
AV15	0.580	0.231
AV16	0.297	0.713
AV17	0.690	0.163
AV18	0.577	0.215

Extraction method: principal component analysis.
Rotation method: varimax with Kaiser normalisation.
Rotation converged in three iterations.

factor and also a low loadings (i.e. 0.20 or smaller) on all other factors is another suitable criteria to apply. However, what you need to remember is the higher the loading the more important that item is to that factor.

So if we look at the rotated solution for our data, we can see that all the items, with the **exception** of AV1, AV2, AV3, AV4, AV6, AV7, AV14 and AV16 load on the first factor above 0.52. The items AV1, AV2, AV3, AV4, AV6, AV7, AV14 and AV16 load on the second factor above 0.56. So, we would suggest that we have two factors on which different items load.

But what about the orthogonal rotation? Well this current analysis, and comparison of the two tables, raises an important point. From factor analytic techniques you are always looking for a clear interpretation. What we mean by a clear interpretation is that all variables should load highly on one factor, and low on all other factors like our example. If you have a number of variables which load on a number of factors, or variables that load on none of the factors then there is something wrong with your extraction techniques. Loading of a number of variables above the criteria of 0.44 and 0.3 across different factors might suggest you have not extracted enough factors, while if a number of variables do not load on any factor then you have extracted too many. (However, don't necessarily concern yourself if one variable loads on two factors. This is known as a cross-loading and may be due to a variable being ambiguous, a bad item (i.e. it should not be there) or genuinely being applicable to both factors.) Also, if you have an interpretation that is consistent with theory, or just makes common sense, then it is useful. However, most of all, you are looking for a clear interpretation with clear loadings.

If we are looking at Table 25.12b then we can see that this is a less clear solution. Items generally load in the same way, with the majority of items loading on factor 1 and the items AV1, AV2, AV3, AV4, AV6, AV7, AV14 and AV16 loading on factor 2. However, one of the items AV5, AV8 and AV12 loads above 0.3 on both factors. This is a less clear interpretation than the previous one where the items are so clearly defined onto one factor (loading highest on one, and low on all other factors). Therefore we would most likely proceed with the oblimin solution, noting in the write-up that we had done both, and the oblimin rotation had produced the clearest solution.

We have a final job to do. We need to give our results some meaning. The next job is to define these factors. The reason we used the word 'define' is that factor analysis does not tell you the nature of the factor, you decide that for yourself. That is, you look at what variables load on a factor and then you give that factor a name. What you call the factor is up to you, it is usually a general term representing the factor. However, the way in which you define the factor is by looking at the loading on the factor. To do this you have to return to the items. We have outlined the oblimin rotation again in Table 25.13.

Table 25.13 Factor analysis output.

(a) Pattern matrix

		1	2
1	I can be spiteful to my friends if they get a better mark than me	−0.072	−0.745
2	In the past I have falsely told other students the wrong exam date, so they would miss the exam	−0.029	−0.799
3	I would hate the student who got the best mark in one of my classes	−0.044	−0.753
4	If I had the opportunity to change other students' exam grades so that mine were the best, I would do it.	−0.003	−0.677
5	I find myself wishing bad things on people that do better than me academically	0.645	−0.125
6	I always tell people I am happy for them when they do better than me in exams*	0.097	−0.601
7	I have thought about spoiling someone's work because it is better than mine	0.017	−0.558
8	I wish bad things on people because they are smarter than me.	0.546	−0.146
9	When other students on my course are praised for their excellent work, it makes me want to wish something bad on that person	0.542	−0.053
10	If my friend got a better mark than me, I would consider tampering with his/her work in some way	0.684	0.121
11	I have mean thoughts towards people who score better than me in exams	0.764	0.031
12	I resent people on my course who excel in their studies.	0.630	−0.124
13	I feel bitter towards those who do better than me on my course.	0.672	0.010
14	I have lied to people on my course to try and hinder their progress so I can get the better mark	0.183	−0.480
15	If I came second best in an piece of work, because my friend got the top mark, I would still be happy*	0.595	−0.048
16	I seek revenge on people who get better grades than me and take away my chances of success.	0.081	−0.722
17	I do not like to help others with their work as it might result in them getting a better mark	0.749	0.074
18	I would consider doing something nasty to somebody who threatened my chances of academic success.	0.597	−0.031

Extraction method: principal component analysis.
Rotation method: oblimin with Kaiser normalisation.
Rotation converged in five iterations.

If we look at the two factors and pick out some of the highest loading factors, you can see the following:

- Factor 1 comprises items such as:
  - I have mean thoughts towards people who score better than me on exams (item 11).
  - I have thought about spoiling someone's work because it is better than mine (item 7).
  - If my friend got a better mark than me, I would consider tampering with his/her work in some way (item 10).
  - I feel bitter towards those who do better than me on my course (item 13).
  - I find myself wishing bad things on people that do better than me academically (item 5).
  - I would consider doing something nasty to somebody who threatened my chances of academic success (item 18).
- Factor 2 comprises items such as:
  - In the past I have falsely told other students the wrong exam date, so they would miss the exam (item 2).
  - I would hate the student who got the best mark in one of my classes (item 3).
  - I can be spiteful to my friends if they get a better mark than me (item 1).
  - I seek revenge on people who get better grades than me and take away my chances of success (item 16).
  - If I had the opportunity to change other students' exam grades so that mine were the best, I would do it (item 4).
  - I do not like to help others with their work as it might result in them getting a better mark (item 7).

We can perhaps see a distinction between these two factors. Items on the second factor seem to represent active and direct vindictiveness towards people, actually being spiteful, seeking revenge, lying to other students, possibly changing their marks. While the items on the first factor seem to be more passive vindictiveness, or much less direct vindictiveness, thinking mean things, feeling resentful, wishing or considering doing something awful to another student. We would argue, and this may be open to debate, that the academic vindictiveness scale contains two factors: direct academic vindictiveness and passive academic vindictiveness. There are some useful things to remember regarding exploratory factor analytic techniques. Factor analysis is a very descriptive procedure. It requires you to describe things, and it is not always a perfect science. This is a criticism of factor analysis as it can be very interpretative. Because of the interpretative nature of factor analysis, it is usually expected that you might perform a number of extractions and different rotation techniques to fully explore possible factors arising from your data. Revisiting extraction techniques, and trying to see whether extracting

Stop and think

? ! ?

Study break

1 In the example above we called the academic vindictiveness direct and passive academic vindictiveness. Do you agree with this interpretation? Looking at these factors, and the item loading, can you think of better names for these factors.
2 Clearly there would be further issues of validity here. What could you do to try to establish validity, not only for these two factors, but to support our distinction between two types of academic vindictiveness: direct academic vindictiveness and passive academic vindictiveness.

more or fewer factors and trying different rotation techniques is acceptable and useful.

However, we would suggest, from our results that this distinction between 'direct' and 'passive' vindictiveness is a possible good way of explaining the data. We would also recommend that rather than computing an overall score, researchers should compute two separate scales' scores, one for direct academic vindictiveness and one for passive academic vindictiveness.

Confirmatory factor analysis

In this section we're going to introduce you to another factor analysis procedure that is used in the literature. We're not going to go into as much detail as exploratory factor analysis because (1) it uses very advanced statistical techniques and (2) demonstration of the procedures require specialist software (e.g. programs called AMOS; LISREL). However, we're going to introduce you to some of the key terms and ideas because the technique is being used more and more within the psychometric literature. Therefore, this section will enable you to read this literature and be able to understand and use it.

Confirmatory factor analysis is a technique that can be used to confirm any findings identified with exploratory factor analysis. So, for example, we can use confirmatory factor analysis to assess whether the two-factor model of academic vindictiveness we found in the above example can be replicated in other samples.

The underlying aim of confirmatory factor analysis is to assess whether any future data collected on the scale fits the explanation provided by the exploratory factor analysis, i.e. does other data collected fit our two-factor model. The key term you will see with confirmatory factor analysis is **goodness of fit**.

What you will then see in many psychometric papers is an attempt to confirm previous models. In Figure 25.9 there is an illustration of the type of diagram you will see presented in a confirmatory factor analysis procedure which shows the suggested structure of the data. In our example you can see that we've connected our 18 academic vindictiveness items onto two main dimensions; direct and passive academic vindictiveness.

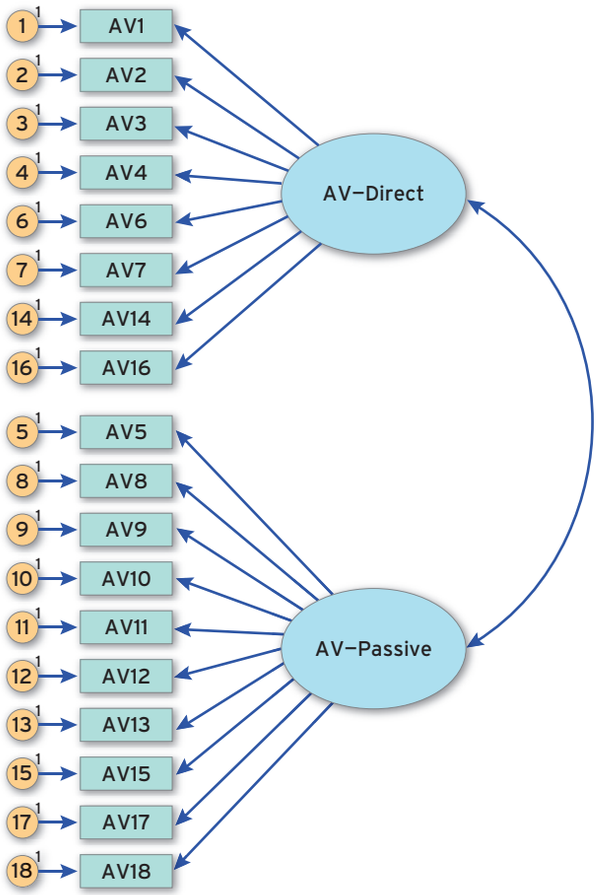


Figure 25.9 Our statistical two-factor explanation of academic vindictiveness.

If we collected some more data on the scale, we could, within a suitable software package, test how well our new data fitted the suggested model outlined above.

We assess the goodness of fit of the data by the use of the goodness of fit statistics. As with any statistical procedure, there are several of these statistics, and what you will find is that authors report a number of these statistics to provide an overall assessment of the goodness of fit. The statistics you will usually see are as follows, and we have provided some suggested criteria by which goodness of fit is assessed.

- chi-square (χ^2)
- the goodness of fit index (GFI)
- normed fit index (NFI)
- comparative fit index (CFI)
- adjusted goodness of fit (AGFI)
- index root-mean-square residual (SRMR)
- the root-mean square error of approximation (RMSEA).

These figures produced by the analysis vary. The chi-square statistic usually provides quite a large number, i.e. around 1000. Analysis for GFI, AGFI, NFI and CFI typically range from 0.00 to 1.00 with higher figures (i.e. those nearer to 1) indicating a better fit. The SRMR and RMSEA produce figures that centre around a low figure, e.g. working towards 0.01, with lower figures indicating better fit. As the chi-squared test is highly sensitive to sample size, it is usually just reported. Authors such as Hu and Bentler (1999) suggest that a good model fit is individually indicated with approximate values of GFI, NFI and CFI above 0.95, and AGFI should also be at least 0.90, SRMR below 0.08, RMSEA below 0.06, should be conventional values for accepting good models.

We subjected some new data we had collected among 300 adults to confirmatory factor analysis. We computed the following goodness of fit from the data:

- chi-square, $\chi^2 = 931.4$
- the goodness of fit index (GFI) = 0.957
- normed fit index (NFI) = 0.950
- comparative fit index (CFI) = 0.988
- adjusted goodness of fit (AGFI) = 0.952
- index root-mean-square residual (SRMR) = 0.038
- the root-mean square error of approximation (RMSEA) = 0.013.

Using our criteria we can see that all our values exceed (in terms of the GFI, NFI, CFI and AGFI) and fall below (SRMR and RMSEA) the criteria, and, therefore, our current two-factor model presents a good explanation of the data.

One final use of confirmatory factor analysis is that it is very useful for comparing different explanations for data. So, for example, if there were two competing explanations for how a scale should be structured, then you could perform confirmatory factor analysis on both interpretations to see which presented the best fit. For example, with our academic vindictiveness scale, we could suggest an alternative idea, that given our earlier findings with the internal reliability analysis, the academic vindictiveness comprised one not two factors.

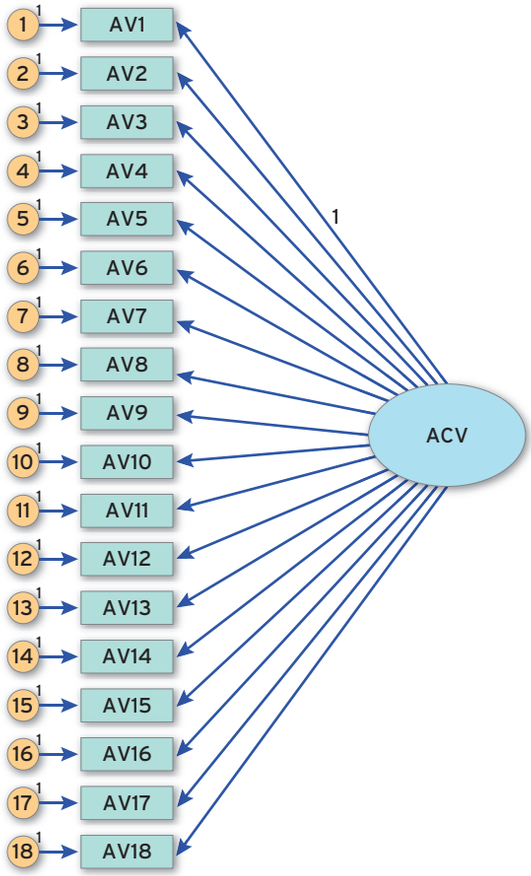


Figure 25.10 Our statistical one-factor explanation of academic vindictiveness.

So, for example, our alternative model would look like Figure 25.10.

We ran a confirmatory analysis on this model and found the following statistics:

- chi-square, $\chi^2 = 826.2$
- the goodness of fit index (GFI) = 0.917
- normed fit index (NFI) = 0.910
- comparative fit index (CFI) = 0.928
- adjusted goodness of fit (AGFI) = 0.882
- index root-mean-square residual (SRMR) = 0.142
- the root-mean square error of approximation (RMSEA) = 0.011.

These fit statistics for our data and model are quite good, in fact, you may see that many researchers would suggest, that though not a good fit, it is certainly reasonable. Nonetheless, the statistics aren't as good as for the statistics for a two-factor model, and therefore we would suggest that there is evidence that our academic vindictiveness scale comprises two factors. You will often see this technique in psychometric papers, with authors comparing different models using these statistics as a basis of comparison, based on either previous findings or

Stop and think

? ! ?

Rasch Analysis and Item Response Theory

A different approach to analysis of psychometric scales was suggested by Georg Rasch in 1960. His original research was concerned with abilities but the theory behind it has now been applied to personality measures as well. The guiding principle behind Rasch analysis was clearly stated by him: 'a person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one'. (Rasch 1960/1980, page 117). In other words when participants complete a psychometric scale they provide us with two sources of information. They tell us how people respond to the items, which is used in reliability and factor analysis, but they also tell us how the participants score on the scale, information, that is not much used in classical item analysis. Rasch's suggests that we should use both pieces of information when we analyse scales.

The basics of the analysis are reasonably straightforward, although the mathematics behind it becomes more complicated. To take the item information first; it is easy to work out the difficulty of each item by using the percentage of the sample of participants who get the answer correct. This can be transformed into the probability of getting the item correct or the odds of getting an item correct. We can also calculate the ability of each participant by taking the percentage of items that they get correct and can then turn this into a probability of that person answering an item correctly. Rasch's theory suggests that the probability of getting an individual item correct is caused by the difference in a person's ability and the item difficulty. To put it simply if a person's ability is higher than a particular item's difficulty then the participant is more likely to get this correct than if it is lower than the item's difficulty. Using this information we can compare the data collected with what we would expect based on calculations of item difficulty and person ability. The closer the results are to the predicted results the better fit the data are to the Rasch model. Like many modern statistical procedures, however, Rasch analysis assumes that the model might be improved by iteration. That is, in this case we can modify are estimates of person and item difficulty so that they converge with the data, but still hold true to a Rasch model. This also allows us to identify which participants are not responding as the Rasch model predicts, and also which items do not fit the model. In both cases, there will be a larger disparity between the results in terms of correct responses and the predicted results.

Rasch analysis is designed to produce unidimensional measures which, therefore, measure only one ability, personality trait or attitude at a time. It is also designed to produce measures in which the difference between participant scores is interval, which is better for statistical analysis.

The first set of fit statistics give us the item reliability which is similar to Cronbach's alpha and the person reliability.

Calculating Fit Statistics

Standardized Residuals N(0, 1) Mean: .00 S.D. : 1.01

Rasch analysis of Maltby

Persons	402 INPUT		402 MEASURED		INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	41.7	18.0	-.79	.30	1.02	.0	1.01	.0
S.D.	12.4	.0	.90	.10	.40	1.1	.41	1.1
REAL RMSE	.32	ADJ.SD	.84	SEPARATION	2.65	Person RELIABILITY		.88
Items	18 INPUT		18 MEASURED		INFIT		OUTFIT	
MEAN	930.3	402.0	.00	.06	1.00	.0	1.01	.1
S.D.	91.3	.0	.25	.00	.12	1.8	.15	2.0
REAL RMSE	.06	ADJ.SD	.25	SEPARATION	4.32	Item RELIABILITY		.95

Figure 25.11

In the example, below we have carried out a Rasch analysis on the data from the academic vindictiveness scale, which has two factors according to the factor analysis. The analysis was carried out on Winsteps (Linacre, 1999/2006) software which is a program specifically designed for Rasch analysis. The first thing to note is that the Rasch solution and the data converged in nine iterations. This would indicate that the data is a reasonable fit to the one dimensional model, given the number of participants and items. The item reliability of the 18 item scale is 0.95 which suggests that all of the items are measuring the same dimension. The person reliability, which tells us whether the participants are behaving in a similar way and is analogous to the item reliability, is 0.88. This suggests that there may be some participants who are not performing like the others (see Figure 25.11).

We can investigate this by looking at person misfit, this indicates the participants who fit the Rasch model less well. In this case, there are five participants who have very poor fit and these are the first five on the SPSS file. Winsteps also indicates the items on which these participants are performing unusually. In this case they score unexpectedly high on items 2, 3, 6 and 16 which are all important in identifying factor 2 (see Figure 25.12).

It is, therefore, possible that the second factor might be attributable to the responses of five participants. If we carry out the factor analysis again omitting these five participants, we find that the second factor now has an eigenvalue of 1.032 which is lower than the criterion for a second factor suggested by parallel analysis. In this case the second factor might be created by five participants who have an odd pattern of responding. It

The table below gives the participants in misfit order with those who fit the Rasch model least well listed first. In this case the first five participants have much worse fit than the others with the first three being equally poor. The outfit MNSQ are all above 3 which is an indication of misfit.

INPUT: 402 Persons 18 Items MEASURED: 402 Persons 18 Items 89 CATS 3.68.0

Person: REAL SEP. : 2.65 REL. : .88 ... Item: REAL SEP. : 4.32 REL. : .95

Person STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT OBS%	EXACT MATCH EXP%	Person
1	50	18	-.20	.22	3.38	5.6	3.49	5.7	A .29	.24	.0	33.0	1
2	50	18	-.20	.22	3.38	5.6	3.49	5.7	B .29	.24	.0	33.0	1
5	50	18	-.20	.22	3.38	5.6	3.49	5.7	C .29	.24	.0	33.0	1
3	51	18	-.15	.22	3.21	5.3	3.34	5.5	D .33	.24	.0	32.8	2
4	49	18	-.25	.23	2.94	4.8	3.08	5.0	E .28	.24	.0	32.9	2

Winsteps also indicates the items on which the participants had the greatest misfit. In this case the first three participants show the same pattern, all with unexpected scores of 5 on items 16, 6, 2 and 3. Participants 4 and 5 also have unexpected scores on 3 of these 4 items.

MOST MISFITTING RESPONSE STRINGS

PERSON	OUTMNSQ	ITEM
		1111 11 1 11
		44865710381625 9723
	high	
1	3.49	A ...5.....55....5
2	3.49	B ...5.....55....5
5	3.49	C ...5.....55....5
3	3.34	D ...5.....5....5
4	3.08	E .....55....5

Figure 25.12

Factor analysis omitting the first five participants, please note the eigenvalue of the second factor.

Total variance explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.947	38.596	38.596	6.947	38.596	38.596
2	1.032	5.734	44.330	1.032	5.734	44.330
3	.893	4.963	49.293			

The Scree plot also suggests one factor (remember the number of points above the elbow).

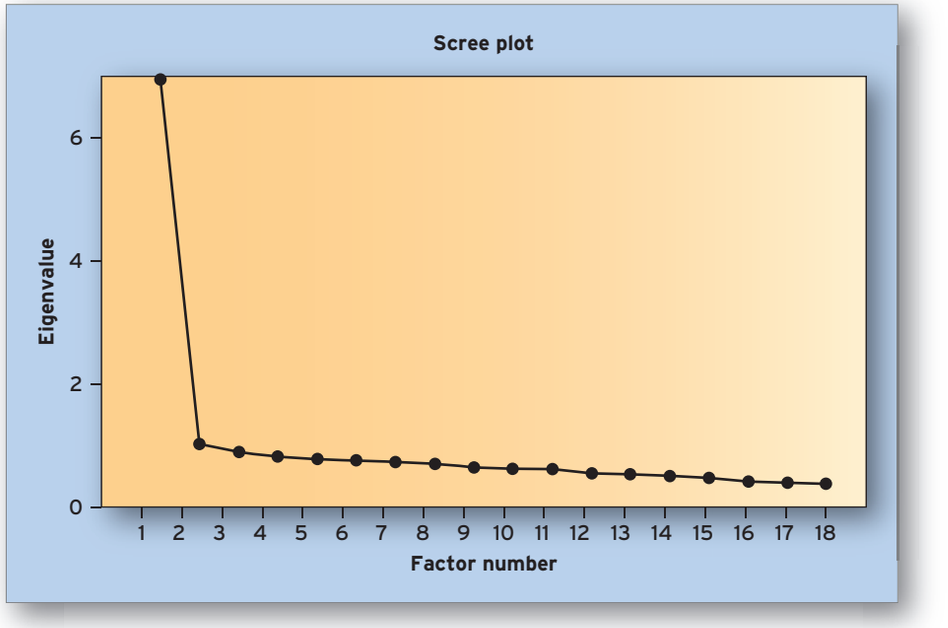


Figure 25.13

is perhaps important to note that some of the items that are aligned on different factors appear very similar; see, for example, items 3 and 12 (see Figure 25.13).

Increasingly there is recognition that psychometric analysis should use more sources of information in its analysis, which has led to the development of a related set of techniques called item response theory (Embretson and Reise, 2000). Rasch/item response theory are often treated as relatively new developments in psychometric testing, however, that the importance of Rasch's discovery was noted over 40 years ago by Loevinger (1965) who stated 'Rasch has derived a truly new approach to psychometric problems . . . Rasch must be credited with an outstanding contribution to one of the two central psychometric problems, the achievement of nonarbitrary measures'.

In conclusion, Rasch/IRT is another statistical tool that can be used by a psychometrician in developing a reliable and valid psychometric tool.

Steven J. Muncer
Durham University

on theoretical perspectives, to find out which models represent the best explanation of the data.

There is a lot more to confirmatory factor analysis but what we've outlined above has given you a brief introduction and provided you with enough information to read some of the psychometric literature.

The International Personality Item Pool and the Higher Education Academy in Psychology practicals webpage

One final point about psychometric tests is that in many instances there is already a measure available to use. We've already mentioned the International Personality Item Pool that contains main personality and individual difference measures (**http://ipip.ori.org/ipip/**). There are other test bases online, but a recent one to appear is at the Higher Education Academy in Psychology practicals webpage (**http://www.psychologypracticals.com/**). This is a searchable library of over 1,500 psychometric tests that are in the public domain. Therefore there is no need to develop new psychometric tests automatically, as there are already many available and these sites are worth checking out.

Final comments

As you can see, psychometric tests are useful in a number of areas of psychology. The main thing to remember about psychometric tests is that it is always important for them show acceptable levels of reliability and validity. Clearly, if you are making decisions about people's education, their livelihood, their treatment for a condition or their illness, you need to be quite confident of your diagnosis.

In this chapter we have introduced you to the main concepts of item writing, reliability, validity, exploratory factor analysis, confirmatory factor analysis and types and uses of psychometric tests. Therefore, you should now understand some of the ideas and criteria surrounding psychometric testing, know what is meant by reliability in psychometric testing, be able to explain what is meant by validity in psychometric testing, see how factor analysis is used in psychometric testing and the types and uses for different psychometric tests.

Summary

- The main types of tests you will come across in the personality, intelligence and individual differences literature are measures of personality, ability, motivation and attitude. Also psychometric tests are used in clinical settings and great care must be taken with their use in assessing people.
- The key to a good psychometric test is writing very good questions. These include making sure the items are clear, not leading, are not embarrassing, don't pose hypothetical questions and include reverse wording items.
- The format of the response scale and instructions are also key to a good psychometric test.
- In psychometric testing, there are two forms of reliability: internal reliability and reliability over time (test–retest reliability). Internal reliability (or consistency) refers to whether all the aspects of the psychometric test are measuring the same thing. Test–retest reliability assesses reliability over time.
- Validity is concerned with whether a test is measuring what we claim it is measuring. Convergent validity is assessed by the extent to which it shows associations with measures that it should be related to. Concurrent validity is when a test shows acceptable correlations with known and accepted standard measures of that construct. Discriminate validity is when the test is *not* related to things that it shouldn't be related to. Face validity is concerned with whether the measure measures what it claims to measure. Predictive validity assesses whether a measure can accurately predict something in the future.
- Getting other people to rate the individual on the items of the questionnaire is a good way of potentially assessing the validity of the questionnaire because, ideally, the ratings given by the participants and the other person should be similar.
- Factor analysis is a multivariate (multiple variables) 'data reduction' statistical technique that allows us to simplify the correlational relationships between a number of variables. There are two forms of factor analysis, exploratory factor analysis and confirmatory factor analysis.
- There are two procedures in exploratory factor analysis; (1) the Extraction of factors which determines how many factors underlie our data and (2) Rotation of factors to determine which items/variables load on each of the extracted factors.
- Confirmatory factor analysis is a technique that can be used to confirm any findings with exploratory factor analysis.

Connecting up

In this chapter we talked about factor analysis. This technique was used extensively in determining the trait approach of personality, such as the five-factor model (see Chapter 7) and crucial in the debate about the nature of intelligence and whether it forms a single factor or not (see Chapter 11). Also throughout the book we discussed different psychometric measures of the constructs we are outlining. In almost all cases many of these measures have established reliability and validity.

Critical thinking

Discussion questions

You might want to try the following two exercises to explore some of the issues that surround psychological testing.

- Researchers in America, such as Professor M. Groening, have developed a new measure of personality, the *Homer Simpson Personality Questionnaire*. Respondents are presented with five items and asked to indicate the extent to which they disagree or agree with each statement on a five-point scale (1 = Disagree strongly, 5 = Agree strongly). Here are the five items of the *Homer Simpson Personality Questionnaire*:
  1. 'It takes two to lie. One to lie and one to listen.'
  2. 'Weaselling out of things is important to learn. It's what separates us from the animals . . . except the weasel.'
  3. 'If something is too hard, give up. The moral is to never try anything.'
  4. 'Hmmmm. Donuts. . . . What can't they do?'
  5. 'Just because I don't care doesn't mean I don't understand.'

  Consider, does this scale have good face validity? What studies could Professor Groening complete to establish or examine the scale's concurrent, discriminate, predictive and construct validity?

- In 2004 and 2005, BBC News reported a number of stories that centred on people's pessimism with the world. In June 2004, a survey by an international polling agency, GlobeScan, suggested Nigerians and Zimbabweans were feeling especially pessimistic about their own countries. In Zimbabwe, just 3 per cent of those asked thought life was getting better. In Nigeria, 75 per cent of people asked thought that the country was heading in the wrong direction, with 66 per cent thinking it was more corrupt than a year ago. In December 2004, Romania's presidential and parliamentary elections on Sunday evoked a mood of pessimism in some of the country's newspapers. In January 2005, there was reported pessimism in France about public sector strikes. In Spain it was pessimism on peace in the Basque country. Also in January 2005, those in the Arab world were thought to have a sense of foreboding as the elections in Iraq fast approached. Throughout 2008 and 2009, Western economic systems are in crisis, with values of shares in stock markets falling, unemployment rising, industries and companies losing money, and most people being effected in some way by the 'credit crunch'. Given such a prevailing pessimism throughout the world's different continents, we are proposing the World Pessimism Trait Scale that measures individual's pessimism about the world. The scale contains five items and asks the extent to which respondents disagree or agree with the statement on a 5-point scale (1 = Disagree strongly, 5 = Agree strongly). Here are the five items of the World Pessimism Scale:
  - I often think that things in the world are never going to get better.
  - I am often of the opinion that the human race is destined towards its own destruction.
  - I am certain that, year after year, the world will become a harder place for everyone to live in.
  - I often wonder what is so wrong with the world these days.
  - The world has so many problems that cannot be solved.

  Consider, does this scale have good face validity? What studies could the researchers do to establish the scale's validity?

- Imagine a health construct you might want to measure. How might you measure this construct? How could you establish reliability and validity (particularly validity) for this construct?

Now imagine a questionnaire isn't suitable to measure this construct and that you have to choose two other methods, an interview and an experimental measure.

– How could you establish reliability and validity of the interview measure?
– How could you establish reliability and validity of the experimental measure?

Now, think about rival ways of measuring the same variable that uses different methods. What are the advantages and disadvantages or each measure, particularly in terms of reliability and validity?

Essay questions

- Discuss how a researcher establishes confidence in the use of a psychometric test.
- Outline the main types of reliability and validity.

Going further

Books

Paul Kline has written a series of readable books about psychometrics. The most recent is Kline, P. (2000). *Psychometrics Primer.* Free Association Books. However, other good books on psychological testing include: Gregory, R.J. (2007), *Psychological Testing: History, Principles and Applications* (5th edn), Harlow: Pearson Education; Anastasi, A. (1988), *Psychological Testing* New York: MacMillan Publishing Company; and Rust, J. and Golombok, S. (2009), *Modern Psychometric: The Science of Psychological Assessment* (3rd edn), London: Routledge.

Within the psychometric literature you will also see references to two other techniques used in psychometric testing; Rasch Model and Item Response Theory (see Stop and think box on page **000**). Item Response Theory is a modern framework for psychometric test construction in which the investigator argues there is a single underlying construct under which all the items reply and each respondent is assumed to have a certain amount of the construct being measured. Rasch Model is a mathematical framework which uses equations to predict the probability of respondents at different levels of skill answering test questions correctly. Like confirmatory factor analysis both these techniques (1) use very advanced statistical techniques and (2) require specialist software to demonstrate these procedures. However, two good books that cover these topics are Wilson, M. (2005), *Constructing Measures; An Item Response Modeling Approach*, New Jersey: Lawrence Erlbaum Associates; and Bond, T.G. and Fox, C.M. (2007), *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2nd edn), New Jersey: Lawrence Erlbaum Associates.

Web links

A good library of psychological tests for public domain use are available at the International Personality Item Pool site (http://ipip.ori.org/ipip/) and a searchable database is at the Higher Education Academy in Psychology practicals webpage (**http://www.psychologypracticals.com/**).

If you want to learn more about confirmatory factor analysis you need a copy of the computer programs AMOS/LISREL. Your department may have these, but they are advanced statistics programs. However, there are some good online guides, for example at **http://www.indiana.edu/~statmath/stat/all/cfa/index.html**.

Film and literature

There are not many films that are clearly illustrative and accurately portray psychometric testing. However, *The Recruit* (2003, directed by Roger Donaldson) a gives a fictional insider's view into the CIA Agency: how trainees are recruited, how they are prepared for the spy game, and what they learn to survive. The film *Spy Game* (2001, directed by Tony Scott) is in a similar vein.

Explore the website accompanying this text at www.pearsoned.co.uk/maltby for further resources to help you with your studies. These include multiple-choice questions, essay questions, weblinks and ideas for advanced reading.

Don't forget the following additional material can be found on the Website (www.pearsoned.co.uk/maltby)