6 Collecting data

Key Issues

Chapter 6 is the second of our procedural chapters and is concerned with gather information – the data which will allow us to answer a research question an issue or test a hypothesis. The first part of the chapter deals with ethical and considers the type of participant you can reasonably expect to take part in search; issues concern vulnerable groups, deception, research involving anir the key features of informed consent. Since it is an expectation that all under research conforms to appropriate ethical standards, we offer specific advice participants should be treated at all stages of the research process. We also p typical consent form which can be used as a template for your own research.

The second part of the chapter deals with questionnaires, a data-gath strument which features prominently in much undergraduate research. We the different ways in which questionnaires can be designed, we discuss diff sponse scales, along with their advantages and disadvantages, and we pro vice on the various stages involved in their development.

The concluding part of this chapter deals with standardised tests; man will consider using existing tests of personality, attitude, ability or some othe ure as part of your study. In this event there is much that needs to be unc about psychological tests and we offer detailed discussion on the evolution their characteristics and how they should be administered and interpret points include:

- carrying out of ethical research
- informed consent
- questionnaire design
- response scales – closed-ended, open-ended, rating scales, Likert scales, tic differential
- the use of standardised tests
- test norms
- test reliability
- test validity

6.1 Carrying out ethical research

It is important to be aware that, before you can inflict a study on an unsuspectin you are bound by certain constraints on your behaviour – especially those co your willing participants. Aside from the implicit moral obligation of any psych

researcher to prevent distress among those individuals who give up their time to help out, there exist a number of guidelines which should always be considered in the design and implementation of any study. Developed over many years of research, and based on broadly accepted moral, behavioural and ethical values, these guidelines have been produced by the various overseeing bodies and are available in full in their many publications, relating both to human and (where appropriate) animal participants.

A summary of the main points of these guidelines is offered below and students should note that, while they are only guidelines, most supervisors would decline to supervise an undergraduate project if they are not adhered to. Moreover, as we appear to be evolving into an increasingly sensitive, not to say litigious society, academic institutions are themselves becoming more cautious about the nature of research carried out in their name. In any event, ignoring these points should be considered only in exceptional circumstances, and with the strongest possible justification since, by and large, they reflect powerful human values and not simply professional ones.

6.1.1 Sampling

Where participants are to be drawn from specific populations (workers in an organisation, patients in a hospital, pupils in primary schools), you should be aware of any possible disruption to normal institutional functioning which your study may cause. It is therefore important that approval and authorisation be sought from appropriate individuals or bodies before your work commences. Indeed, some institutions (e.g., hospital boards) require research proposals to pass their own form of ethics committee before approval is given. Furthermore, any reporting on the findings of a study must include details of all procedures used to obtain participants, to ensure the consent of participants and to seek the approval of relevant bodies.

6.1.2 Apparatus

Apparatus refers to any instrument, device or questionnaire which is used to aid the collection of data. In the case of standard equipment, such as a tachistoscope (a device for back-projecting images onto an enclosed screen for pre-determined durations), it is important that all operations are fully understood and the regulations governing use are fully adhered to. In the case of standard questionnaires and psychometric tests, the instructions for administration and scoring must be followed. Further, no such instrument should be used without a thorough awareness of norms, limitations, applications, reliability and validity studies, such that in no way can participants be disadvantaged by your lack of familiarity with the manual. (This issue is covered in detail later in this chapter.)

In the case of non-standard equipment, details of any unusual features should be included, insofar as they have a bearing on the design of the study; otherwise it is enough to state that, for instance, in a maze-running experiment, 'a maze was constructed comprising an equal number of left and right hand turns'. In the case of non-standard questionnaires, a full copy, with rationale for each element, must also be included as part of the report. Undergraduates often fail to do this, to their cost, since it is often the case that findings will be interpreted in terms of the procedures used to develop and score items. The reason for so much detail is not simply replicability, although this is often important, but also to ensure that ethical standards are maintained. If a researcher is unable to provide such depth of information or, worse, is unwilling to do so, then the research is clearly suspect.

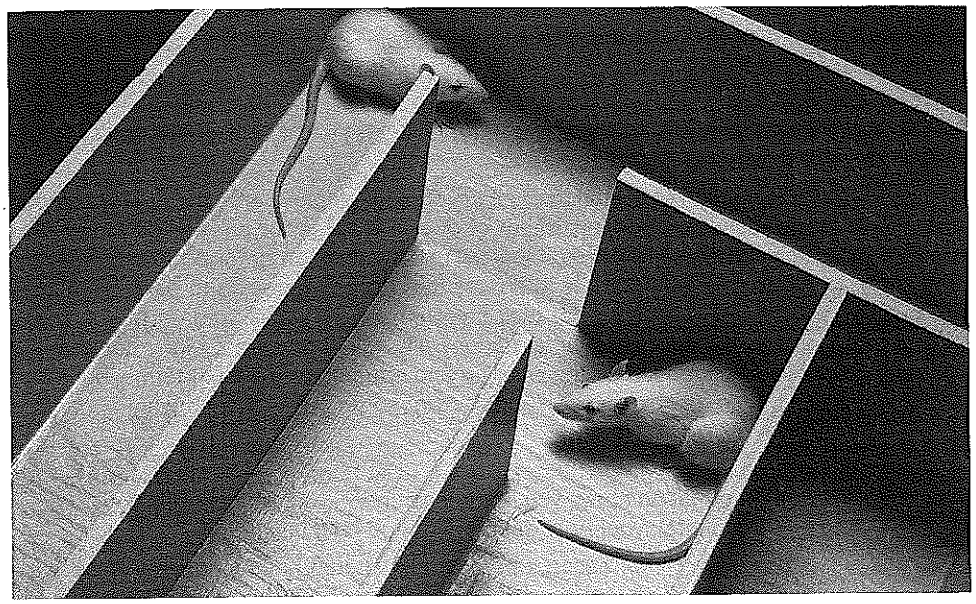


Plate 6.1 Various mazes can be constructed for laboratory rats to explore, differing in complexity.
Source: Science Photo Library/Will & Deni McIntyre

6.1.3 Procedural issues

Participation in any psychological study is normally voluntary and you must ensure that participants are aware of this, and that they are able to withdraw *at any time* during a study, without prejudice (i.e., without any fear of sanction and with an assurance of no negative consequences of withdrawal), a right which extends even beyond the data-collection stage. At no time must you coerce or use deceit (which is not in itself a part of the study) to obtain co-operation. In the event that some type of deception forms part of the experimental manipulation, you must ensure that participants are fully debriefed (i.e., told exactly what has happened and why).

If a situation is encountered in which participants are not able to provide **informed consent** themselves, steps must be taken to protect the individual: if people with neural damage, young children, or individuals exhibiting other forms of cognitive disorder are to be approached as participants, consent must be obtained from those who have their interests at heart. These could be relatives, parents, carers or medical staff. And even then, unless there is sound reason for pursuing a particular research interest, such individuals should not be used in a study. Increasingly, academic departments are now requiring participants to sign a consent form, stating that they fully understand the purpose of the study in which they are participating and that they are aware of their rights, both legally and morally. Box 6.1 offers an example of a typical consent form which might reasonably be used as a template for much undergraduate research.

If the data collected in a study are to be kept confidential (as they must be), you must take steps to ensure that not only is this so, but that it is seen to be so, especially by your participants – people who have willingly agreed to participate in your research in the belief that they will be treated fairly and with respect. This is especially true if the data might be considered private and personal to participants. Again, reported research is now expected to include details of how the confidentiality of responses has been ensured.

BOX

Checklist . . .

What should be in a consent form

It is now an expectation of all psychology departments that, before their students embark on testing or administration of questionnaires for a research study, they obtain the informed consent of participants, usually through an appropriate consent form. The purpose of such a form is to ensure that anyone agreeing to participate in research does so knowing precisely what their involvement will entail, and on the understanding that their rights to confidentiality and anonymity will be ensured. Many departments will have their own preferences for the wording and structure of consent forms, but for those of you who will be expected to devise their own, some guidelines on what should go into a consent form follow:

- Information about you, the researcher – your name, your status (e.g., a level 1 undergraduate), your host department and host institution.
- Contact information – participants should have a contact address or phone number in the event that they have concerns about the research, or about their rights. Normally the contact individual would be your supervisor or a member of the departmental ethics committee which approved the research.
- Information about the nature of the research – a description of what the study is about and a statement of aims should be given. The issue of clarity and level of language is important here since many people will not understand technical or complex terms and the age and level of sophistication of the target population should always be taken into consideration. Explaining that a study is investigating aspects of field dependence/field independence would certainly baffle most people, whereas explaining that the research is looking at the different ways in which people see the world might go further in informing participants what the study is about.
- The nature of the participants' experience – that they will be taking part in an experiment which will involve looking at a series of images, that they will be discussing issues in a group situation, that they will be completing a questionnaire on a number of issues.
- Sensitive issues – if any element of the experience might raise sensitive issues for the participant this should be made clear.
- The time demands of the study – you must be honest; if you feel that explaining the study will involve staring at scores of images on a computer screen for two hours would put participants off, this is no justification for claiming otherwise. Perhaps the design can be changed!
- Information on what will be done with the data – who will have access to it (and who will not).
- An assurance of the participant's anonymity and the confidentiality of all data collected as part of the research.
- Confirmation that the participant's involvement in the research is voluntary and that he or she may withdraw at any stage.
- An agreement check box or space for initialisation – if the participant is required to sign as an indication of consent this will conflict with the assurance of anonymity.
- Contact or information details – if a study is likely to raise issues over which participants might have concerns, contact details of individuals or organisations who can provide further information should be offered. For example, if you intend to explore

people's understanding of phobias, you would be expected to include information or helpline numbers for the National Phobics Society, or a local health education centre.

A sample consent form:

Thank you for agreeing to take part in this study.

My name is (*name*) and I am (*status; department; university or college*).

This study is investigating (*nature and aims of study*).

It will involve (*nature of participant's experience; the time demands of the study; sensitive issues*).

Please note that your participation in this study is entirely anonymous and voluntary, and that you may withdraw at any point. All the information gathered during the course of this study will remain confidential and will be seen only by (*state who will have access*).

If you have any concerns about this study or would like more information, please contact (*contact person and contact number*).

If, having read and understood all of the above, you agree to participate in this study, please place a tick in the box below.

☐

Optional:

If you have any concerns about issues raised during this study, please contact any of the information lines below:

(*contact details*)

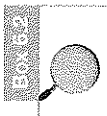
As a general rule, participants should never be placed under undue stress. However, should induced stress form a part of a study, advance preparations must be made in terms of controlling such stress, and for preventing or dealing with possible catastrophic consequences (both physiological and emotional). Generally speaking no undergraduate would be permitted to conduct a study along these lines nor, unless in exceptional circumstances, would an experienced researcher. The days of Milgram, in which participants were placed under great personal stress in the false belief that they were inflicting pain on others, have gone! (Note: references for both the British Psychological Society [BPS] and American Psychological Association [APA] sets of ethical guidelines are provided at the end of this book. These should be consulted prior to the commencement of any undergraduate study.)

6.1.4 General

All of the above are merely guidelines to enable you to conduct yourself and your research in an ethical, humane and fair manner. They should not be regarded as constraints, rather as a series of reminders that when you carry out a piece of research you are dealing, not with abstract sources of data, but with real people who have rights of privacy, sympathy, and expectations of fairness of treatment to which all of us are entitled.

While most departments, and certainly most supervisors, are keen to encourage initiative and creativity among their students, it is likely that ethical considerations will increasingly be an overriding factor in determining the type of research which is granted approval. Vulnerable groups of any kind (children, hospital inmates, prison inmates) are

unlikely to feature among 'approved' populations for study. Topics which may prove stressful, disturbing or anxiety provoking to participants will not generally be approved. The use of standardised tests which might force vulnerable participants to confront difficult issues would not be allowed. In addition, students are now required to present to all potential participants a consent form in which their rights (both moral and legal) are clearly defined (see Box 6.1). There will inevitably be exceptions to these general guidelines – some undergraduates will have legitimate access to special groups (perhaps through working in some counselling capacity with a socially vulnerable group), some may be allowed to participate in ongoing departmental research and there may be other circumstances in which a normally prohibited topic may be approved. By and large, though, there will be real restrictions on what an undergraduate will be allowed to tackle. The point to be remembered is that, in most instances, undergraduate research is for demonstration and assessment purposes, and for the gaining of experience. Consequently, deviating from any of the guidelines governing ethical research will rarely be justified at this level. Box 6.2 offers a review of ethical guidelines.



A Closer Look At . . .

Ethical guidelines

All psychological research today must adhere to a set of ethical guidelines designed to protect the rights and preserve the dignity of participants, while at the same time ensuring the safety of the researcher. In its most recent publication, the BPS working party on ethical practices has produced a set of minimum guidelines (BPS, 2004) which it believes should represent best practice in psychological research. In addition, every psychology department today has in place its own recommendations governing research at all levels, from undergraduate to postgraduate. Below we offer a checklist of questions you should ask of your own research. Answers to these questions will determine whether a study will be conducted in an ethical manner, with due concern for the welfare of participants.

A. Matters of openness and the rights of the individual

1. Will participants be informed that their involvement in a study is entirely voluntary?
2. Will it be explained that participants are free to withdraw from the study at any time, both during its conduct and after its completion, without prejudice?
3. Will it be made clear to participants that their contribution to any part of the study will be totally confidential and that it will be impossible for any individuals to be identified by any party not directly involved in the research?
4. In the event that individual contributions will not be anonymous (perhaps the identity of participants must be retained in order to match different scales) how can participants be assured of confidentiality beyond the requirements of the research design?
5. Will it be made clear what participation in the research will be like? In particular will participants be informed about time demands, about the nature of any activities required of them or about the type of questions they might be asked as part of the study? This is especially important in cases where test or interview items are designed to explore personal or sensitive issues.

6. If a study is questionnaire based, will it be made clear that participants need not answer questions they do not want to answer?
7. Will participants be invited to complete a consent form in which all the points mentioned above are fully explained?

The point of the above guidelines is to ensure that involvement in any study is based on the principle of informed consent; that is, with the full knowledge of what the study is about, with an understanding of what the experience will be like, and with an awareness of the steps taken to protect rights and dignity.

B. Matters concerning special and vulnerable groups

8. In the event that participants are unable to give informed consent (the sample may include very young children, or individuals who have difficulty in understanding or communicating), what steps have been taken to ensure the informed consent of those responsible for their welfare? (These might be the parents or teachers of children, or the carers of other disadvantaged groups.)
9. If a study will involve members of a school, an element of the health service, a community organisation or a commercial business, what steps have been taken to obtain authorised approval to carry out the particular research?
10. If a sample to be used in research might be described as vulnerable – insofar as participants might respond negatively to questioning or to items on a measurement scale what steps have been taken to deal with potentially catastrophic situations? (Confronting a group of phobic individuals with probing questions about their fears could produce a range of negative reactions from feelings of discomfort to uncontrollable panic attacks.)

Research involving special or vulnerable groups raises a variety of important issues for the researcher. Whenever children are involved in research our society is increasingly concerned about protecting their wellbeing and currently, aside from the need to obtain permission from parents and teachers and school officials, the researcher will require to be vetted by an appropriate criminal records authority – the Criminal Records Bureau (CRB) in England and Wales, and the Scottish Criminal Records Office (SCRO) in Scotland. Most departments now have in place procedures for interacting with the appropriate authority. Moreover, depending on the nature of the research, some studies will require approval by the ethics committee of the relevant education authority.

Research in the health service carries with it additional problems in that now all studies must be scrutinised by an appropriate ethics committee, arranged through the newly formed (at the time of going to press) Central Office for Research Ethics (COREC). Moreover, research within any type of organisation risks disruption of care, productivity and social structures and the researcher will be required to take steps to minimise such disruption.

With vulnerable groups especial care is required to protect participants from all forms of harm. If there is a danger of individuals becoming distressed during the course of a study, safeguards must be set in place to provide counselling or therapy and the contact numbers of appropriate helplines should be made available as an additional support for the research participants.

C. Deception

11. Is there an intention to mislead participants about the purpose, or about any part of the study?

12. If deception features in a study, will participants be fully debriefed on completion of their involvement? Will they be informed that they have a right to withdraw from the study following debriefing?

The use of deception in research is a controversial one, in which the need to research a particular issue must be balanced against all the points made previously about informed consent and the rights of the individual. Within the context of undergraduate research, though, the issue will almost never arise since, with research at this level, a need to know will never be sufficient to outweigh the need for openness.

D. Research with animals

Non-human research is unlikely to feature at all at undergraduate level and for this reason we merely note that (rightly) a number of rigorous and legally binding safeguards are in place governing the care and welfare of all animals used in research. The BPS offers advice for anyone planning research of this type.

6.2 Using questionnaires in research

6.2.1 Questionnaire design – how to get information

By this stage in our research, we have hopefully decided on the issue we are going to explore, we know whom we are going to use and we have decided on the key components of our hypotheses. We also know what kind of information we want from our participants, the precise data needed to test our hypotheses and to ultimately explore the research issue in question. So how are we going to get this information?

Some research designs will use standardised instruments to generate information (health questionnaires, stress measures, personality tests, etc.); others will rely on an outcome measure of a laboratory experiment (reaction times, frequency of correct responses to stimuli, or changes in decision times). Many designs, though, will require custom-made procedures to gather information, as researchers devise questionnaires to assess attitudes to numerous issues, to obtain information on what people do in various social situations, to measure opinion on a wide range of social and political issues or to explore the distribution of different categories of person in the population. Gathering information of this sort might appear, on the face of it, straightforward – they are all simple question-and-answer scenarios, whether they involve interviews or questionnaires. The reality is somewhat more complicated in that the development of a 'good' questionnaire is not just a skill but almost an art in itself. The following sections attempt to make the process a little easier, explaining the pitfalls and offering solutions.

Probably the simplest rule of information gathering is 'if you want to know something, ask', and, by and large, this is the most useful rule to follow when designing a questionnaire or interview schedule. Just ask your participants to tell you what you want to know. Most people are honest, disingenuous and, once they have agreed to participate in a study, usually willing and co-operative. Unfortunately, a common perception of psychology is one of a somewhat sneaky profession, relying on methods of deception and misdirection for its information. Even among students of the discipline, there is a view that participants have to be tricked in some way into giving honest, objective responses and,

unfortunately, such a view will only continue to encourage the sense of suspicion and mistrust directed at elements of the profession by outsiders. Such a lamentable state of affairs has its origins in the type of research allowed before guidelines were established by the psychological societies and various educational and medical associations which govern the activities of their members. Contemporary researchers now frown on the needless use of deceptive techniques (see the earlier section on ethics). In most instances – with certain qualifiers – the direct approach is best: if you want to know something, ask. The qualifiers are important, though.

The general rule of *ask and ye shall be answered* holds true most of the time. But sometimes the nature of response can be influenced by who does the asking and how the asking is done. Consider the following question:

> Have you ever, at any time in your life, committed a crime?

This is an apparently simple question, if asked by a social researcher guaranteeing absolute confidentiality. Imagine the nature of response if the same question were asked by a serving police officer conducting research into criminal behaviour among undergraduates. So the *who* of a question is important and what every researcher must ask him or herself is: 'will the fact that *I* am asking a particular question affect the response?'

The other major qualifier is the *how* of a question. The example above (Have you ever, at any time in your life, committed a crime?) can only generate either a Yes or No response. Modifying the question to . . .

> What crimes, no matter how small or insignificant, have you ever committed in your life?

. . . is likely to produce a very different class of response. Aside from the fact that this could be described as a leading question (it assumes people do commit crime) the asker has no control over the type and quantity of response. Anything from a 'How dare you . . .' to a two-page list of guilt-ridden confession is possible. And so, the *how* of a question is an important consideration. The following section describes the most common ways of asking questions and the kinds of responses each produces.

6.2.2 Types of scale

There are two broad types of question available to the researcher: one in which the researcher controls the nature of the response and one in which the participant is free to respond in any way. Both have their uses and their disadvantages, and the decision as to which method to apply is one the researcher must make in terms of the context of the research and the quality of the information required.

Closed-ended questions

Closed-ended questions occur where the possible range of responses is pre-determined by the tester. (The opportunity for free response is closed to the person answering.) There are many forms of this type of item, the simplest of which allows for answers on a dichotomous scale.

Table 6.1 Responses of 200 psychology students to an attitude question item: 'Do you like the book?'

Response	*n*	%
Yes	120	60
No	80	40

Dichotomous-category scaled items

Dichotomous scales are used for questions offering only two answer choices.

Do you like the book?

Yes No

The data questions like this produce are in the grand tradition of survey techniques. Participants respond by selecting one or another of the nominally scaled categories and sample data can be presented simply, by referring to the numbers, proportions or percentages of participants who selected each category. Table 6.1 and Figure 6.1 demonstrate the economy and elegance of this approach.

Table 6.1 and Figure 6.1 both display an at-a-glance summary of the data, and the Yes/No distinction represents one of the simplest – and most common – item formats available to researchers. However, care must be taken not to confuse simplicity with impoverishment. True, the Yes/No response options provide only limited information; but when other variables are introduced, the basic dichotomous distinction suddenly becomes quite sophisticated.

Consider the above example when we wish to further analyse the Yes/No choice in terms of the student's gender, or age group, or study options; suddenly we have more than straightforward descriptive data – we can begin to make comparisons and to make inferences. We have now moved, and quite painlessly at that, to a point where the number-devouring statistician begins to take an interest (see Figure 6.2).

The dichotomous example is only one of a variety of closed-response formats. There is no reason why we should stick to just two possible responses when, with most questions we might want to ask, there are invariably several types of response possible.

Multiple-category scaled items

Multiple category scales are used for questions offering three or more choices for the respondent.

Figure 6.1 Responses of 200 psychology students to an attitude question: 'Do you like the book?'

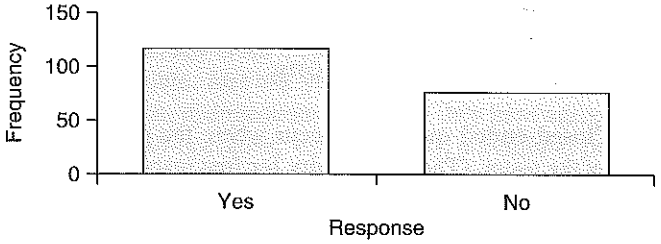
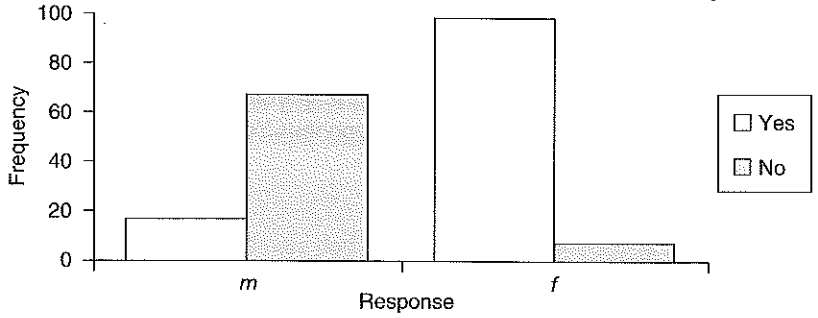


Figure 6.2 Responses of males (*m*) and females (*f*) to a questionnaire item: 'Do you like the book?'

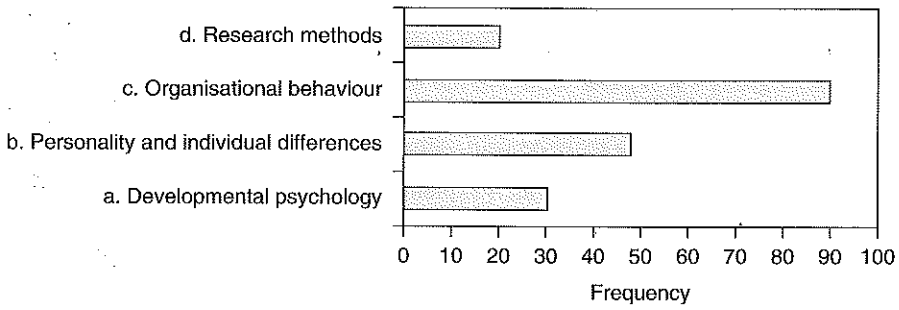


During the current semester, which of the following modules have you enjoyed most?

a. developmental psychology ☐
b. personality and individual differences ☐
c. organisational behaviour ☐
d. research methods ☐

Just as with the previous example, the response categories are independent of one another: that is, there is no relationship of magnitude or order between any one category and another, only of difference – the essence indeed of all nominal scales. Similarly, the display of this type of data is equally straightforward, as Figure 6.3 demonstrates:

Figure 6.3 Frequency of response (*f*) to the question: 'During the current semester which of the following modules have you enjoyed most?'



Rating scales

Rating scales are scales which rate some attribute from negative to positive, low to high, weak to strong.

Moving along the continuum of sophistication, but still retaining control of how participants can respond, are scaled items. Instead of requiring participants to choose from a number of response categories which differ in type from one another (as in the dichotomous and multiple examples above), we focus on just one category and require participants to indicate their strength of feeling on the issue. In the 'do you like this book' example, participants could give only the following responses: Yes (they do) or No (they don't).

Interesting as such responses are, they nonetheless obscure the range of feeling within each category; that is, one Yes respondent might be transcendentally ecstatic about the textbook, whereas another might merely be expressing a 'yeah, it's OK' attitude. This kind of internal distinction is lost in fixed-category items, but is accessible in rating scales in which, at its most basic level, nominally scaled responses are transformed into ordinally scaled ones. Consider the re-structuring of the previous Yes/No item.

I am enjoying this book:

not at all ... tremendously

1 2 3 4 5

Not only does this provide a more detailed picture of how strength of feeling varies across an issue, it also moves the relevant information away from the descriptive and towards the more traditionally quantitative. What this means is that, while in the previous examples participants differed in the type of response they made, now participants differ in terms of, at the very least, the order and even magnitude of response. It also moves the analysis of data towards a format with which many quantitative researchers feel more comfortable: we now have measures of central tendency (average) and variability to play with. In other words, we have **parametric data**.

However, the presentation of more sophisticated data like this should be no less straightforward than for the earlier examples, provided we are familiar with the concepts of sample statistics. Table 6.2 and Figures 6.4 and 6.5 demonstrate this:

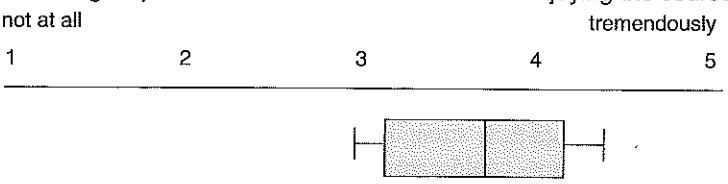
Table 6.2 Mean and standard deviation of response to an attitude question.

	(mean)	(standard deviation)
I am enjoying the course:	3.75	0.64

The mean provides a measure of central tendency (average) of the responses on the issue and the SD (standard deviation), a measure of how much, on average, scores varied around this value. Part 4 explains these concepts in greater detail. See Figure 6.4.

In Figure 6.4 the median, or middle value, indicated by the dark bar in the middle of the rectangle (some statistical packages represent this as an asterisk *), is approximately 3.7; the interquartile range (the range of scores from the lower quarter to the upper quarter of the distribution of scores) as indicated by the upper and lower limits of the enclosed rectangle, is approximately 3.2 to 4.2; and the overall range of responses, shown by the

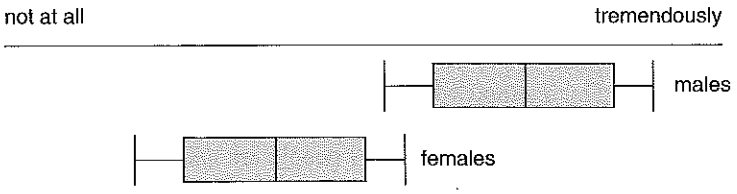
Figure 6.4 Boxplot illustrating responses on an attitudinal item: 'I am enjoying the course'



upper and lower 'whiskers', is approximately 3.0 to 4.4. This particular method of descriptive illustration is termed a boxplot and is further explained in Part 4.

As with our categorical examples, we can of course make our analysis more sophisticated with the inclusion of additional independent profile or participant variables. For instance, we can compare the different genders on the same scale, or different seminar groups, or whatever. The example shown in Figure 6.5 demonstrates this:

Figure 6.5 Responses of male and female participants to an attitudinal item: 'I am enjoying the course'



A special variant of scaled items comes to us from Likert (1932). In this approach, participants are asked to provide their level of agreement with a statement. Usually (though not always) corresponding numerical values are absent and appended only later when the researcher converts response categories to their quantitative equivalents. For example, in the following illustration, participants choose one of the agree/disagree categories. The choice made will ultimately place their attitude on some kind of (assumed) linear scale and means that standard deviations and all kinds of comparisons can be produced in the time-honoured manner of parametric statistics, even though by rights this is actually an ordinal scale, with the numbers referring to categories rather than points on a continuum. Many researchers, especially undergraduate ones, choose to ignore this, however, and continue to treat these types of data as if they were interval and hence susceptible to parametric analysis. There is debate on the issue and arguments range from the purists who would gladly have their students shot for treating category data as if they were continuous, to the pragmatists who take the line that 'well, if it's going to show something, why not?' Definitely a case for checking with your supervisor. The de Vaus reference in our suggested further reading section also provides useful background material on different types of scale. (Note: if these statistical terms are foreign to you, the next part of this book introduces basic statistical concepts.)

Likert scale

The **Likert scale** is a response scale where the respondent indicates the amount of agreement/disagreement with an issue. It is usual to construct Likert scales with an odd number of response categories (typically five), allowing for a neutral central category.

Psychologists are nice people.

strongly agree	agree	neutral	disagree	strongly disagree
(+2)	(+1)	(0)	(−1)	(−2)
5	4	3	2	1

Normally, with this type of item, respondents are presented with only the worded response options (*strongly agree, agree*, etc.). The numerical scales are shown to indicate how the researcher might transform actual responses to numerical scale values. Two such transformations are shown here, one in which the middle value is given as a zero on a positive-to-negative scale, and the other in which this value is shown as 3, on a 1-to-5 scale. If the researcher is going to pretend that the scale represents a numerical continuum of values (as in an interval scale) the variant with a zero point is probably the most useful, certainly the most intuitive, since the zero point on the scale represents an absence of opinion or view. Adopting a similar line with the 1-to-5 scale will lead to problems of interpretation, since values of 3 on a series of items might be intuitively interpreted as reflecting a stronger attitude than values of 2 or 1, when in fact scores of 3 represent an absence of opinion or attitude strength. Far better to treat the numbers as categories on an ordinal scale and avoid confusion altogether.

This method whereby actual scale values are obscured can have its advantages. A problem with asking participants to choose a numerical value indicating a particular view or attitude is that sometimes people are unclear as to how their feelings can convert to a number; or they may be reluctant to select extreme values, or they may be unsure of how one scale value differs from the next. Replacing numbers with choice categories (as in the Likert scale) will sometimes alleviate this problem, in addition to making items more user-friendly. A development of this approach in which participants are indirectly placed on some scaled position is the semantic differential, shown below:

Semantic differential

With **semantic differential** the respondent rates an issue on a number of bipolar categories. The choice of pole can indicate intensity of feeling on the issue.

Research projects are: (choose one of each pair of response options)	
good	bad
easy	difficult
useful	worthless
challenging	problematic
interesting	boring

In its simplest form, items at one pole can be given a positive value, and items at the other pole a negative value. A simple arithmetic count of both positive and negative choices can produce an overall score which will be indicative of the general attitude towards an issue.

Visual analogue scales

Expressing attitudes or feelings is often difficult since such things don't always translate easily into a number. The semantic differential described above is one way of dealing with this. Another method is the **visual analogue scale**. Typically this is a horizontal line, 100 millimetres in length and anchored at each end by one extreme of an issue. Such scales can be used to measure like or dislike for something, and for measuring constructs such as fear and anxiety. They have also been used extensively in pain research.

Please indicate with a cross on the line how much pain you are in at the moment.

_____X_____

no pain at all extreme pain

Since such scales are of a standard length (100 mm), an ordinary ruler can be used following administration to convert a response to a numerical value on a 10-point (10-centimetre) scale, as shown:

_____X_____

1	2	3	4	5	6	7	8	9	10

Open-ended questions

With **open-ended questions** the individual is free to offer any response. The researcher relinquishes control of how the respondent can behave. In fact, open questions are not really questions at all. Rather, they are scenarios or situations created artificially in which respondents are encouraged to air their views or feelings. In this respect they have much in common with what are termed **projective techniques**, as used in a clinical setting, or for the measurement of personality variables. Given free choice, we all tend to project some part of ourselves onto a situation, and this is the principle behind the open-ended question. What we say or write in response to an unrestricted item is an expression of our feelings, values and attitudes.

Unstructured items

Unstructured items are questionnaire items which allow the subject to answer in an unlimited fashion.

What do you think of the book so far? _____

Note that in items of this type the researcher does in fact have some control over how much of a response an individual can make, although the content will still remain unpredictable. Leaving the item as it stands invites a very brief response, possibly only a single word. Allowing several blank lines, or even a page, might encourage a much fuller response.

Word association

Using **word association**, the individual is asked to respond with the first thing that comes to mind on presentation of the cue word.

What do you think of when you hear the following?

research project exam statistics

Sentence completion

In **sentence completion**, the participant is offered an incomplete sentence and asked to continue in his or her own words.

I chose to study psychology because _____.

Note that the comment about controlling the length of response (as encountered in the 'What do you think of the book so far?' example earlier) is relevant here.

The great advantage in not restricting responses is that the full variety of human thoughts and feelings is available to the researcher. The disadvantage is that it is available in a totally unstructured and uncontrolled form. Responses to any of the above can range from a few words to hours of introspective rambling. Accordingly, the approach lies more at the qualitative end of the research dimension (in terms of the type of information generated), but for quantitative researchers the data produced by open-ended questioning are more problematic: given the wide range of possible responses, considerable judgement is required and each participant's response must be inspected for common themes or threads of thought which offer an insight into the unique nature of the individual.

As a general principle, unless the components of an issue and the range of possible responses to a question are well understood, a study would normally be piloted using some form of open-ended enquiry. This would identify the type and range of responses likely, any sources of ambiguity, possible response bias among the respondents, overlap among items and so on. Only then would the more direct, closed-type items be used. The rule in closed-type items is simple: you will get only the information you ask for, so you had better have a good idea of what you are likely to get!

6.3 Special techniques for information gathering

6.3.1 Focus groups

The previous section on questionnaire design ended with the sage advice that if you don't ask the questions, you won't get the answers. But how do you know which questions to ask? Or to what extent a particular issue is important or relevant to your sample group?

It would seem then that, even before we consider our very first questionnaire item, we should know in advance what we are looking for. In many cases, this will have been established following an appropriate literature review: issues would have been identified at this stage and hypotheses developed. However, there are certain situations in which a research issue, or its key components (and hence a workable hypothesis), cannot readily be identified in this way; many elements of attitudinal and behavioural research are simply not accessible via the standard route and have to be explored in a more direct manner. One particular method of doing this, originating and widely used in the marketing sphere, is the focus group.

One of the key functions of the marketing process is to find out how people are likely to respond to new products, publicity campaigns and marketing information. If this is done well, we, the consumers, respond positively when a product is released on the market

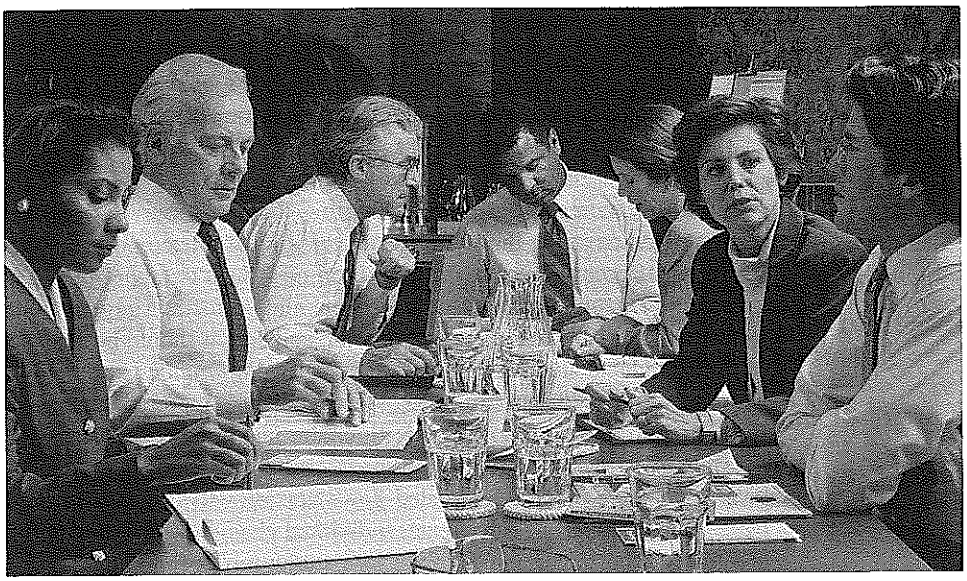


Plate 6.2 A focus group.
Source: Getty/Daniel Bosler

and (so the manufacturers hope) go out and buy it. Poorly done, the result is a marketing disaster and possible bankruptcy. The attempt by Ford to launch America's first small, economical automobile in a culture of gas-guzzling monsters was an abysmal failure, and even today comedians make jokes about their parents being the first in the street to buy an Edsel. Similarly, the attempt by British Leyland in the 1970s to persuade us that an Austin Allegro with a square steering wheel was a must, ended in fiasco. (Oddly enough, the square-wheeled Allegro is now something of a collector's item.)

These examples represent classic marketing disasters – public opinion was seriously misjudged and important informational cues missed. By and large, however, marketing strategists tend to get it right (or right enough) by relying on a variety of techniques to gauge opinion, evaluate campaigns and judge consumer response. One of the mainstays of the approach is the focus group.

Much as the term suggests, focus groups are essentially discussion groups comprising either randomly or carefully selected panels of individuals brought together to discuss, or focus on, specific issues. Discussion can be free-flowing or controlled, but it is always under the guidance of a moderator, or facilitator, whose role is to maintain the focus of the group's attention on the issue, and to further probe or develop important themes. As an exploratory instrument, focus groups are superb sources of information, allowing a skilled researcher excellent insight into the values, beliefs, fears and aspirations which comprise most attitudes. Not surprisingly then, the approach has become an important tool in recent years in social, and especially political, research. Opinion, though, is divided as to both the value of the focus group and the procedures involved. In some quarters, the focused approach is seen as a preparatory procedure only, a way of refining ideas, getting to grips with the scope of a particular issue or developing a theory sufficiently to generate hypotheses (as intimated in our discussion of open-ended questions). For others, the focus group is an end in itself, with the information generated providing the essential and necessary data for a study, as discussed in Part 6. And certainly, given the vast amount of

potential information generated in this format – hours of audio tape, pages of transcriptions and so on – its appeal as a qualitative research tool is obvious. However, it is in the first context that focus groups are probably of most use to undergraduate researchers, providing as it can a way of coming closer to an issue, of developing a greater understanding of the components of attitudes and of identifying key or relevant issues which will ultimately form the basis of a questionnaire.

With its increasing application within the behavioural sciences, several guidelines have evolved covering the use of focus groups and related procedural issues: how to sample, optimal group sizes, the composition of groups (whether or not members should be strangers or people known to each other) and so forth. There are views on the roles of moderators, on how to collect information and how it should be analysed. For the majority of undergraduate studies, though, a focus group *approach* is probably of more value than adopting a full procedure. In most cases a student is merely interested in identifying or refining important issues so that questionnaire instruments can be designed which will ask the right questions and provide sufficient coverage of an issue to deal appropriately with a given research topic. Such a scaled-down version will still require planning – participants have to be recruited to participate; individual characteristics have to be identified if these are going to be important variables in a study; topics for discussion have to be prepared, along with procedures for guiding or focusing discussion, dealing with awkward individuals and recording data. Equally important, especially if the research is dealing with a sensitive topic, is the thought which must be given to protecting vulnerable members of the group and on dealing with upset or catastrophe. Generally speaking, untrained undergraduates would be discouraged from using a focus approach to explore highly emotive or disturbing issues without a great deal of supervision and forward planning. Indeed, most research ethics committees would reject an undergraduate proposal to explore feelings and attitudes towards sensitive issues when participants of the focus group were themselves the victims of, for example, abuse or assault. Approval of this type of study might be granted only if supervisors were present at each focus session and experience of intervention could be guaranteed.

By and large, many undergraduates tend to use this procedure badly, calling brief, informal discussion meetings with relatively few individuals, failing to adequately control discussion and failing to record data in any systematic way. Often this process is seen merely as a precursor to the more important business of interview schedule or questionnaire design; students forget that they will ultimately have to justify every issue covered, and every item contained in any instrument. This can be done only if the issues have been properly explored and understood in advance. In Part 6, the use of focus groups in qualitative research is discussed in more detail.

6.3.2 Pilot research

Most of us feel that by the time we reach the stage of implementing our research design, we have worked out everything to the last detail: we have completed our literature review and therefore know how other practitioners have fared in their research; we have identified all potential sources of bias; and we have used an appropriate procedure to focus on the key issues and develop a foolproof questionnaire. However, complacency at this stage is to admit to a poor regard for the vagaries of human nature – misunderstanding instructions, misperceiving the researcher's intent, refusal to co-operate and so on are all events

which can ruin the best conceived study. The solution of course is to pilot your method: try it out on a small sample of the population you will eventually be working with. This is the only way to refine the elements of a design, to identify questionnaire items which are misleading, confusing or offensive. Such pilot work need not be extensive – indeed in designs where participant pools are limited, pilot studies must be constrained – but they can be thorough: a survey or questionnaire administered to a small sub-set of our sample, in addition to some kind of focused interview, can be useful in identifying limitations and areas of improvement. Mistakes at this stage can be easily remedied; identifying flaws only after a major study has been implemented is hugely wasteful, not to mention demoralising.

6.3.3 Using standardised instruments in research

Many studies will make use of standard scales, existing questionnaires or psychometric tests as part of their data-gathering procedure, either as devices for assigning individuals to different conditions, or as a key source of data, as in an outcome measure. For example, an existing stress inventory could be used to assign people to the categories of either stressed or unstressed in a study of job burnout in modern organisations, categories which will comprise the various elements of an independent variable. Alternatively, standard instruments could provide us with our dependent measure, as in a wellbeing questionnaire measuring the impact of unemployment in particular regions, or an established job satisfaction instrument assessing responses to supervisors differing on androgeny scores in an academic context. By and large, using existing scales as part of a study can make life a little easier for the researcher: normally, before a test can be released on to the market, it must demonstrate that it is fit for its purpose; a great deal of preparatory work has invariably gone into the design and construction of any psychometric instrument. This provides the researcher with a useful measurement or classification tool without the need for the lengthy process required in developing a new instrument from scratch. However, using an existing test correctly requires familiarity with the general principles underlying measurement and scaling, in addition to an understanding of the essential characteristics of the particular test or questionnaire itself. Such a level of competence can be attained only through many years of experience with measurement scales; furthermore, the major publishers and distributors of tests have for some years provided training courses in various aspects of assessment while, more recently, the BPS has introduced an accreditation scheme whereby potential test users are obliged to undertake specific training before they are regarded as qualified to use particular tests. This is an important consideration, since failure to understand how a test has developed, or what specific responses mean, or a failure to prevent bias in administering or scoring, can lead to misinterpretation of scores and misleading – or even damaging – information being fed back to the testee. Equally dangerous is the potential for exploitation when tests are used by unqualified or unscrupulous people and it is largely for these reasons that the BPS introduced its scheme.

The observant reader might have realised by now that most undergraduates are unlikely to have either the experience or training to include existing scales as part of their research repertoire with any personal competence. The assumption is that it is the competence of the supervisor and the department to which they belong which allows the use of such instruments, not the students, although in all cases the users will be expected to familiarise themselves thoroughly with whichever test they will be using.

6.4 What you need to know about psychological tests

While it would be nice to test every aspect of some issue under investigation, such that we could explore a group or culture's entire history, or measure every aspect of an individual's life as it relates to the issue we are interested in, this is clearly impractical. (Any such test would be not only unwieldy, but would probably take a lifetime to administer.) What established instruments actually do is study a small but carefully chosen sample of some topic, issue or behaviour, in the hope that we can generalise from the specific to the global, in much the same way that a survey, while its main interest is the entire population, can only ever explore a small section of that population. Hence, a vocabulary test cannot address every word in a person's repertoire; rather it deals with a sample of what the individual knows.

To be of any use in predicting or describing what people do in general, this sample of items must be **representative** of the overall area, both in terms of type and number of items. An arithmetic test using only five items, for instance, or only items on multiplication, would be a poor test of arithmetic skill. A good test, on the other hand, would have to include items on addition, subtraction, multiplication and division. Furthermore, we might be unhappy if such a test omitted items involving fraction or decimal calculations and we might also expect some measure of computational abilities such as dealing with square root, power or factorial calculations. The underlying point here is that it would be impossible to develop a good, representative test of any aspect of human endeavour unless the composition of that behaviour had been fully considered in advance – in our arithmetic example we could develop a sound test only if we had a thorough understanding of the scope and composition of arithmetic skill to begin with.

While we expect test designers to demonstrate a sound knowledge of their particular area, it is equally important that we, as ultimate users of tests, also understand a good deal about the issues we are exploring. How else could we judge whether or not a test was a good one for our purposes, or, given a number of similar tests, how would we know which was the most appropriate?

6.4.1 Standardisation

Even the best measurement scale in the world will be wasted unless we can ensure that scores reflect the subject we are interested in, as opposed to some other factor. Notorious 'other factors' which can affect apparent performance on a test are instructions given to participants, their level of anxiety about the test, motivational factors, methods of collecting data and scoring procedures – nothing less, in fact, than the extraneous variables discussed in Chapter 2. Unless every individual who completes a test does so under identical, standardised conditions, any observed effects might simply reflect **procedural variations** rather than actual differences in behaviour. Box 6.3 illustrates the point.

Fortunately, test constructors are well aware of this issue and are able to employ several procedures to reduce the effects of administration variability. The Eysenck Personality Questionnaire, or EPQ (Eysenck & Eysenck, 1975), a favourite among psychology undergraduate researchers, is a good example (see Figure 6.6). It's a pre-printed test with restricted response categories and instructions clearly printed on every copy. Even scoring has been taken out of the hands (or judgement) of the administrator, being achieved via

Box 6.3

A Closer Look At . . .

The standardisation problem

Instructions given to a group of university students prior to the administration of a standard intelligence test:

Instruction A

The test you are about to complete is one of the most advanced tests of intellectual functioning yet devised. Your scores on this test will be considered by a team of experts and their decision will partially determine whether or not you are allowed to enter the honours stream next session. It is therefore important that you do well.

Instruction B

I'm afraid today's video presentation is cancelled due to the technician's inability to remove the cling film wrapping from the tape. For want of something better to do, we've found this intelligence test – it's not an especially good test but it might be a bit of fun and it will give you something to do.

Each set of instructions is in its own way inappropriate (and also unethical) in that each actively cues respondents to approach the test in a particular way. It would be unsurprising to obtain two completely different sets of scores not measuring intelligence, but more likely motivation or – especially among those receiving Instruction A – test anxiety.

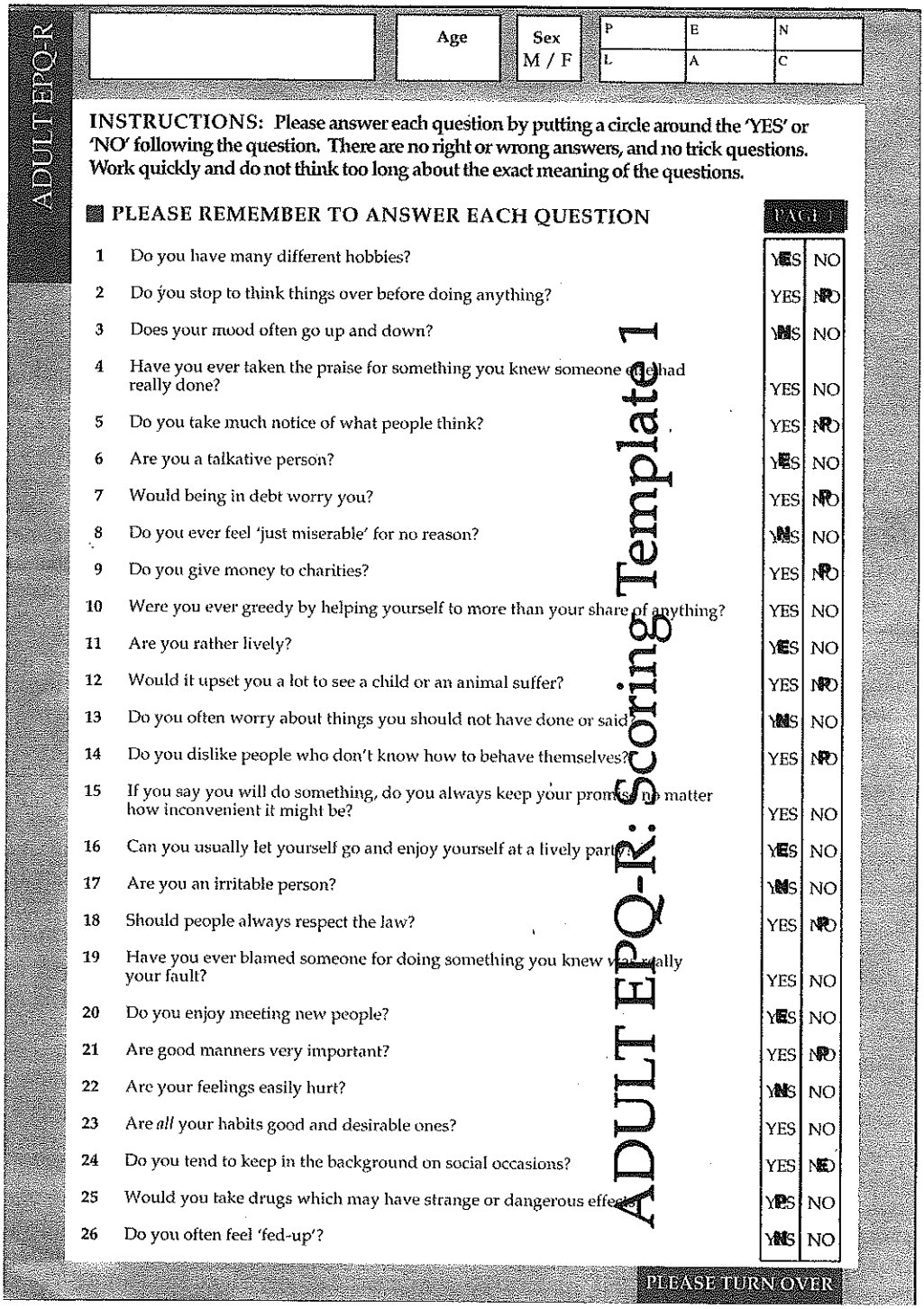
standard scoring stencils. Finally, interpretation of individual profiles can be guided by reference to printed norms, a procedure with which all test users must become familiar.

6.4.2 Norms

Contrary to popular belief, there is no pre-determined pass or fail level in most tests. It would be nonsense to talk of passing a stress or personality test, for instance, although the notion might seem less bizarre if we are dealing with something like arithmetic for which some form of pass levels can realistically be set. (Most students will be aware that the various exams they sit, which test their knowledge of or competence in particular subjects, have clearly defined pass and fail levels.) In fact, outside of such measures of achievement, in the majority of tests, individual scores are compared against other scores which have previously been collated by the test designer. This comparison function is obtained by first administering the test to a large, representative sample of those with whom the test will subsequently be used (the **standardisation sample**). This provides us with a **norm**, which is simply a measure, or series of measures, indicating how people typically (or normally) perform on this test.

Norms can take various forms, although usually they comprise a measure of average performance (being the arithmetic average, or mean of the scores of all the participants in the standardisation sample), and a measure of the extent to which scores tend to vary above and below this average (given as a standard deviation; see Part 4). The point of

Figure 6.6 An extract from the Eysenck Personality Questionnaire, showing the scoring template, superimposed. (Copyright © 1991 H. J. Eysenck and S. B. G. Eysenck. Reproduced by permission of Hodder & Stoughton.)



ADULT EPQ-R

	Age	Sex M / F	P	E	N
			L	A	C

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.

PLEASE REMEMBER TO ANSWER EACH QUESTION

PAGE 1

1	Do you have many different hobbies?	YES	NO
2	Do you stop to think things over before doing anything?	YES	NO
3	Does your mood often go up and down?	YES	NO
4	Have you ever taken the praise for something you knew someone else had really done?	YES	NO
5	Do you take much notice of what people think?	YES	NO
6	Are you a talkative person?	YES	NO
7	Would being in debt worry you?	YES	NO
8	Do you ever feel 'just miserable' for no reason?	YES	NO
9	Do you give money to charities?	YES	NO
10	Were you ever greedy by helping yourself to more than your share of anything?	YES	NO
11	Are you rather lively?	YES	NO
12	Would it upset you a lot to see a child or an animal suffer?	YES	NO
13	Do you often worry about things you should not have done or said?	YES	NO
14	Do you dislike people who don't know how to behave themselves?	YES	NO
15	If you say you will do something, do you always keep your promise no matter how inconvenient it might be?	YES	NO
16	Can you usually let yourself go and enjoy yourself at a lively party?	YES	NO
17	Are you an irritable person?	YES	NO
18	Should people always respect the law?	YES	NO
19	Have you ever blamed someone for doing something you knew was really your fault?	YES	NO
20	Do you enjoy meeting new people?	YES	NO
21	Are good manners very important?	YES	NO
22	Are your feelings easily hurt?	YES	NO
23	Are *all* your habits good and desirable ones?	YES	NO
24	Do you tend to keep in the background on social occasions?	YES	NO
25	Would you take drugs which may have strange or dangerous effects?	YES	NO
26	Do you often feel 'fed-up'?	YES	NO

ADULT EPQ-R: Scoring Template 1

PLEASE TURN OVER

these measures is that, provided the people who tried out the test during these initial stages are representative of the broader population, what is true for the sample should be true for everyone. This is a crucial point: for statistical purposes a standardisation sample of 600 or so participants would be fine, providing what we are measuring is pretty stable in the population. If, however, a test is being designed to measure some trait which is known (or suspected) to be influenced by many factors (it can vary with age, sex, social class, occupation, etc.), then the sample would need to be much larger to allow the different sources of variability to be reasonably well represented. In the development of the Eysenck Personality Inventory, or EPI (Eysenck & Eysenck, 1964), the forerunner of the EPQ, the test developers used a standardisation sample of more than 5,000 participants. With a sample this large, they were able to explore scores on the extraversion and neuroticism scales by age, sex and some 47 different occupational groups, a diversification which would not have been possible with smaller numbers of participants. The point of this protracted discussion on norms and samples is that, if a test is to be used as part of a student project (or indeed for any research purpose), it is important to be aware of how relevant the norms are to the group being studied. A standardisation sample which is small, or which does not allow for sources of variation, might be of limited value in describing the larger population.

Returning to the EPI, the large standardisation sample allows us not only to state that the average extraversion score for all males in the sample was 13.19 and for females, 12.60, but also that the average for male students was 13.80 with female students 13.49. Similar information is available for many other occupational sub-groups and for different age groups. See Box 6.4 which shows an extract from age norms taken from the later EPQ-R (Eysenck & Eysenck, 1991).

Providing an average measure of performance on a test is a common method of presenting norms, but there are others. More typical, and more informative in some ways, are norms which are expressed as **percentiles** – a measure of the percentage of the standardisation sample which scored at or below a particular level. Hence, if a test manual informs us that a score of 35 on an abstract reasoning test is at the 50th percentile, then we know that 50% of the standardisation sample scored 35 or less; if we are told that a score of 59 lies at the 95th percentile, then we know that 95% of the standardisation sample scored 59 or less, and so on (see Figure 6.8).

6.4.3 Test reliability

Every measuring instrument, if it is to be of any use, must demonstrate a number of important qualities. The first of these is that it must be sensitive to the character of whichever variable – be it some physical property or a social phenomenon – it is measuring, and accurately detect any changes which might occur. Equally important, it must not indicate change where no change has taken place. This may sound strange, but consider an everyday measuring instrument, such as a ruler. If we measure the height of a table on a Monday and obtain a measure of 1 metre, but on Tuesday obtain a height of 1.3 metres, there are two things which can have occurred. The first is that the table (for some bizarre reason) has changed its dimensions in the course of a day and is now 0.3 metre taller than it was on Monday, a change which has been accurately detected by our ruler. Alternatively, the table has not changed; rather it is the ruler which has changed – possibly some temperature-sensitive metal was used in its manufacture and the thing actually shrank overnight. Before the advent of plastic-coated materials, traditional fabric tape measures

A Closer Look At . . .

Test norms

In the development of most psychological tests, one of the tasks of the designer is to determine typical score profiles for the population for whom the test is intended. Conventionally this takes the form of a measure of average (usually the arithmetic mean) and a measure of how much, on average, people tend to vary about this mean (given as a standard deviation). There are other methods for indicating typicality – we can determine what percentage of respondents are likely to score at, or below, particular levels of a test, in which case our typical scores take the form of percentiles, or percentile ranks. (We might find that 80% of respondents tend to score on or below a given score on some test, which would make this score the 80th percentile. People performing at a higher level might score on the 90th percentile, and so on. See Part 4.) All such attempts to demonstrate typicality are part of the process of establishing norms: they are measures of how people normally respond to a particular test. Some tests are offered with only the most general of norms (e.g., we will be given typical scores for broad groups only, such as those for males and females). Other tests which have received more thorough development might have gender norms subdivided by age category. And some might even provide norms by gender, age, occupational category and so on. Table 6.3, extracted from the EPQ-R, shows extraversion and neuroticism scores for two different age groups.

The interpretation of normative data like these requires some basic statistical understanding, and Chapter 7 (Part 4) provides a good introduction to the various descriptive measures presented in Table 6.3. Figure 6.7 illustrates the distribution of extraversion scores for males in the 16–20 years age group, using the normative data. The mean is shown as the score obtaining the highest frequency, and the variation around this mean is given as standard deviation units. Specifically, we see that the average extraversion score for this sample

Table 6.3 Age norms (mean and SD) for extraversion and neuroticism. Adapted from the *Manual of the Eysenck Personality Scales (EPS Adult)* (Eysenck & Eysenck, 1996, Table 2).

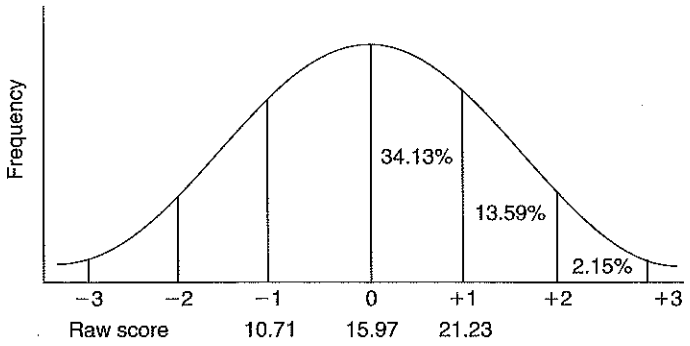
Males

	Extraversion			Neuroticism	
Age (years)	***n***	**Mean**	**SD**	**Mean**	**SD**
16–20	108	15.97	5.26	11.12	5.68
41–50	55	11.91	5.09	11.22	5.95

Females

	Extraversion			Neuroticism	
Age (years)	***n***	**Mean**	**SD**	**Mean**	**SD**
16–20	161	15.47	4.99	14.03	4.85
41–50	50	12.36	4.95	10.94	5.92

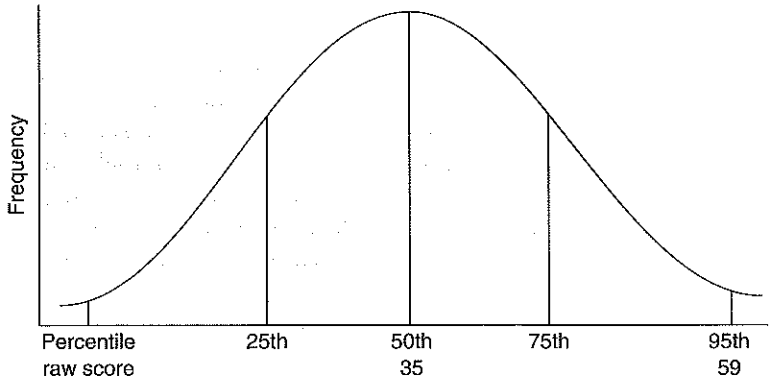
Figure 6.7 Mean and standard deviations of extraversion scores for males in the 16–20 age group.



is 15.97, with respondents varying, on average, 5.26 points above and below this score. Another way of expressing this, given the characteristics of a normal distribution, is that 68.26% of this sample scored somewhere between 10.71 and 21.23 on this particular test.

Note that the mean score given in Table 6.3 for extraversion relates only to one particular age group for male respondents on this particular version of the EPQ scale. Average extraversion scores for other age groups differ from this, and there are clear gender differences as well. Moreover, different versions of these scales show quite marked differences in mean scores. In the EPQ-R Short Scale, for instance, the male mean for the 16–20 age group is 8.16, based as it is on a smaller number of items. The point is that there is a danger – commonly found among undergraduates – of latching onto numerical values as if they are absolute measures of particular constructs. It would not be the first time that a student has tried to interpret an individual score on a personality test in the belief that average performance could be expressed as a given value (such as 15.97), while forgetting or being unaware that norms will vary for particular age groups, for males and females and for whichever form of a test has been used.

Figure 6.8 Scores and percentile ranks on an abstract reasoning test.



were notoriously unreliable in wet conditions and it was the wise surveyor who refused to work in the rain. The point of this example is that, with a standardised test, we always try to ensure that the first case is the typical one: when a change is signalled it must be the aspect of the psychological environment which is changing and not the instrument itself. In other words, a test must demonstrate consistency, or **reliability**. There are a number of ways in which this might be done:

Test-retest reliability

The simplest and most direct method for demonstrating that a test is behaving consistently is to measure the same thing twice – **test-retest reliability**. The same test is administered to the same participants and the two scores compared, using a form of correlation procedure. Obvious problems here are practice effects, which could actually produce a poor comparison between two presentations of the same test simply because our testees were getting better at the task. This could lead us to wrongly assume the instrument is not reliable when in fact the opposite is the case – the test has been sensitive to the change in behaviour. Alternatively, if participants recall their original responses and attempt to reproduce them, an artificially high correspondence between the two measures might occur, demonstrating more the reliability of participants' memories than the stability of the test. Only tests not affected by repetition can be assessed in this way, or alternatively, a sufficient time interval between pairs of testings can dilute the effects of both practice and memory. Of course, leaving a lengthy time interval between testings allows for the intervention of other variables which might affect performance on the task, which merely serves to indicate that demonstrating test reliability will never be completely straightforward.

Alternate form reliability

When there is a danger of straightforward repetition producing a misleading reliability measure – as in the memory issue – two different forms of the test can be given to the same participants and the scores correlated, referred to as **alternate form reliability**. Care should be taken to ensure that the two forms of the test are truly equivalent and of course practice can still influence performance.

There is a method for assessing reliability without the need for two repetitions of the same or parallel forms of the same test – **split-half reliability**. In this variant, a test is administered once only, the items split in half and the two halves correlated. Sometimes a straight split is made, with the first half of a test compared with the second half, while at others times the split is in terms of odd-numbered items versus even-numbered ones. However, this method is not so much a measure of temporal reliability (consistency over time) as of internal consistency (where all the items themselves are measuring in the same direction), and should not be regarded as an alternative to either of the previous approaches; this provides a check of how reliable the individual items are. If a good comparison is achieved, the items are deemed to be consistent and the test largely reliable (insofar as all the items are measuring the same construct). There are of course a number of problems in attempting to demonstrate reliability, with practice and memory effects having already been discussed. Equally important are procedural and administration factors – any variations across successive repetitions of a test will adversely affect reliability measures. Nor is the split-half approach without its problems, because the strength of the comparison may vary according to

Box 6.5 A Closer Look At . . .

Reliability

It is an important characteristic of all psychological tests that, whichever aspect of the individual they measure, they do so consistently. This means that test scores must be closely linked to what is being measured: if an aspect of behaviour remains constant in real life, the test should reflect this stability. If the aspect of behaviour changes, test scores must also change. This is known as reliability and all well-developed tests ought to be able to demonstrate this. Moreover, measures of test reliability and explanation of how these measures have been obtained must be readily available to the test user. The test manual which accompanies most tests will normally offer such information, and the extract from the EPQ-R manual, shown in Table 6.4, is a good example.

The extract in Table 6.4 is typical of the information found in a good test manual, in support of the reliability of a scale or scales. Other information which one would expect to find in a manual would be the type of reliability demonstrated (test-retest or alternate form), the size of the sample used to investigate test reliability, the precise nature of the reliability coefficient and a full citation of the authors of the research. In the current example, in addition to the tabulated information of which Table 6.4 is an extract, we are informed elsewhere in the manual that the reliabilities shown are test-retest reliabilities, with one month between administrations; that the sample comprised 109 males and 120 females; and that the reliability coefficients are given as alpha coefficients (Eysenck & Eysenck, 1996, pp. 14–15). The many studies investigating the reliability of the scales are also cited in full with brief descriptions of the findings in each case.

Table 6.4 Test-retest reliability coefficients for males and females on the extraversion (*E*) and neuroticism (*N*) scales of the EPQ-R. Adapted from the *Manual of the Eysenck Personality Scales (EPS Adult)* (Eysenck and Eysenck, 1996, p. 19, Table 5).

	E	*N*
Males	0.83	0.76
Females	0.89	0.81

how the split is made. There are statistical ways of dealing with reliability problems, and the actual correlation between two testings is not based on the conventional calculation most undergraduates are familiar with, but rather on a development which attempts to counterbalance the deficiencies of the reliability procedure in general. The manuals which accompany most tests ought to provide details of the procedures used to establish reliability, together with the associated **reliability coefficients**. See Box 6.5.

Test validity

The next crucial quality that a test should possess is **test validity**. This is simply an expression of the extent to which a test is actually measuring what it is supposed to be measuring, although the methods available to demonstrate this quality are often far from

simple. In terms of test characteristics validity is possibly even more important than reliability. After all, reliability only tells us that, whatever the test is measuring, it is doing so consistently. It doesn't necessarily inform us about what the test is actually measuring, or how good a job it is doing of measuring it. If a test is demonstrably valid, though, we know it is doing what it claims to do. Unfortunately, in the historical development of testing some of the early forms of intelligence test, for instance, were subsequently shown to be invalid when it was observed that they could not be completed successfully without a conventional Western educational background – they had more to do with scholastic aptitude than with what people regard as pure intelligence (whatever that might be). This is an aspect of validity, and this example also offers an idea of the scope of the problem. Given that many tests attempt to measure the traits which *underlie* behaviour (creativity, personality, intelligence, etc.), the problem of validity becomes a difficult one. Having said this, a number of methods are available which go some way towards demonstrating the fitness of particular tests, and the different types of validity are discussed in the following sections.

Content validity

If a test is to demonstrate **content validity** then the content of the test must accurately and adequately reflect the content of the phenomenon under investigation. In the case of an opinion or attitude measure for example, the content of the test should be based on a thorough understanding of the attitude, related views, associated measures and likely values expressed; knowledge tests should be a fair representation of the topics which comprise the information base. Your exam at the end of a typical methods module, for instance, should be a good expression of the course of study, reflecting topic diversity, issues raised and recommended readings given. If it does not meet the terms of these criteria students have every right to complain about the lack of content validity present in their exam (the content of the test failed to reflect the content of the course). However, many such tests can become overloaded with items which lend themselves to objective testing. While it is easy enough to explore an individual's familiarity with information aspects of an issue, how do you measure something like critical appraisal? And, as previously mentioned, early intelligence tests primarily comprised items on academic skill rather than abstract reasoning, creativity and the like. Achievement tests in particular will invariably be examined in terms of their content validity.

Face validity

Face validity is often confused with the previous test feature since, as with content validity, the concern is with the appearance of a test. However, face validity is not a true indication of validity, being concerned only with what a test appears to measure, and having little to do with what is actually being measured. Nevertheless, this represents a useful feature of any test because there will be a relationship between how people view a test and their willingness to participate; if we see an instrument as childish, irrelevant or even insulting, we will certainly not give it our full attention. Box 6.6 illustrates this issue.

In some cases, especially in the area of opinion studies, it is not always possible to ensure face validity: in a situation where a participant may not respond honestly if the true purpose of the test is known, we may be tempted to disguise part of the test content. This of course must be done with extreme caution and only with considerable justification. It should be understood that this approach actively deceives participants and adherence to ethical guidelines (as published by the various research governing bodies such as the BPS and APA) must be ensured. See Box 6.7 for an illustration of the problem, and also the section on questionnaire design in which this issue is further explored.

Box 6.6

How To . . .

Ensure face validity

> Orange ice lollies cost 50p each, raspberry lollies 35p and lemon ones 40p. If a schoolboy has £1.60 to spend in his tuck shop, and he wants to buy each of his three friends a different flavour, which flavour of ice lolly can he then buy for himself?

The above problem would be a good (valid) item in a test of general arithmetic reasoning. However, if the item appeared in a test designed for trainee electrical engineers, the response would more likely be derisory laughter than the correct answer (which is raspberry lollies, for the computationally challenged). For this particular group the item would not have face validity and some re-wording would be in order:

> Assume 1-millimetre copper cable costs £35 per 100 metre drum, 1.5 mm cable £40 per drum and 2.5 mm cable £50 per drum. If a project buyer needs electrical cable of each size and has £160 to spend, of which diameter cable can he buy 200 metres and stay within his budget?

This is the same item as the previous one except, for this group, it now has face validity. (By the way, the correct answer is 1-millimetre cable, in case you haven't got the idea yet.)

Criterion related validity

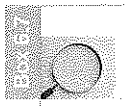
When a test is developed to diagnose something about an individual's present circumstances, or to predict something about a person's future, validation can sometimes be achieved by comparing test scores to some other indicator (or criterion) of what the test is trying to measure. An occupational selection test, for example, can be checked against later job performance (which is actually how such tests are validated); a neuroticism test can be checked against medical records, or friends' ratings of behaviour; a scholastic achievement test can be checked against assignment ratings; and so on. Within this general procedure of relating test scores to some other criterion, there is a particular condition concerning the temporal relationship between the test and its criterion measure (i.e., when we actually obtain this validating information). This relationship is determined by the nature of the test itself – whether or not it is measuring something about an individual's current circumstances, or whether it is predicting something about the future.

Concurrent validity

The criterion against which scores are to be checked is obtained at the same time for **concurrent validity**. This is the type of proof which is necessary when a test is assessing some aspect of a current condition (as in a diagnostic test).

Predictive validity

In a test which predicts something (as with aptitude tests), a follow-up study is carried out to test the strength of the prediction, called **predictive validity**. Most selection tests, as used by industrial or occupational psychologists, will be obliged to demonstrate criterion



A Closer Look At . . .

Concealing the true nature of research

Armed with a limited budget and charged with identifying the 20% of pensioners most in need of additional subsistence payments, you devise a questionnaire to determine differing levels of deprivation.

Please indicate on the scale below how adequate your pension is in meeting your individual needs.

1	2	3	4	5
totally inadequate	inadequate	acceptable	adequate	more than adequate

(Note: the wording used for the response categories would be chosen to reflect the type of question and the nature of the respondent.)

If other items on the questionnaire are like this, approximately 99% of our sample will fall into our most needy category – simply because it is obvious what the questionnaire is about, and what the consequences of particular responses will be. And human nature being what it is . . .

Modifying the appearance of items, however, might provide a subtler if more indirect route to the information you are looking for – for example, if items appear to be measuring more general behaviour than obvious levels of deprivation:

Please indicate how often you eat a hot meal in the course of a week?

1	2	3	4	5
never	rarely	sometimes	often	every day

Or

Approximately how much do you spend on fuel in the course of a week?

Or

On average, how many times do you go shopping during the week?

Items like this allow us to infer certain things about respondents, indirectly, and it could be argued that for some types of information this form of indirect questioning is the only way of ensuring an honest or unbiased response. However, an approach of this type is potentially deceptive insofar as the purpose behind items is obscured, and runs counter to the spirit of openness which forms the basis of current ethical principles. One might adopt the line that the ultimate aim of this particular study is to improve the lot of as many individuals as possible, and moral rectitude will overcome any niggling doubts about deception. However, suppose the same type of study were used in order to reduce benefits to individuals deemed well enough off not to need so much financial support. Therein lies a dilemma for the researcher – is the need for information so great that the means by which we obtain it are always justified?

related predictive validity, since everything from an interview to a job sample test is predicting something about job candidates. If a selection test is described as having predictive validity, then during its design phase it might have been administered to job candidates as part of a general selection and recruitment process. Subsequent assessment of individuals actually hired by the company would be compared with the predictions made on the original test, and if a good match is obtained, the test is declared valid. Only then would it be developed for future use in the selection process.

A problem here is criterion contamination, in which the independent measure can become contaminated by knowledge of the test results and therefore ceases to be truly independent.

"He looks sick, what do you think?"

"Yeah, now that you mention it . . . "

This is a tricky problem to overcome in many cases since it is common for the individual responsible for an original assessment to be the same person involved in subsequent evaluations. The only way round this is to ensure that independent assessments are truly independent – follow-up measures should ideally be taken by individuals who have no detailed knowledge of previous evaluations.

Construct validity

Construct validity indicates the extent to which a test measures some theoretical construct or concept, such as intelligence, creativity or personality. Not surprisingly this is the most difficult type of validity to demonstrate, since the concepts being measured – as the name suggests – are really theoretical entities whose existence is inferred by observation of related activities. Consequently, validation of such concepts is also indirect: measurements of activities which are believed to be related to, expressions of, or caused by, some underlying factor.

Age differentiation is one such indirect method of validation. If a trait is expected to vary with age, scores on the test should reflect this variability. For instance, if the understanding of certain concepts (e.g., prejudice) is part of a developmental process, we should be able to observe this by comparing older children with younger ones. If people become more conservative as they get older, measures of attitudes towards many issues should differ between a middle-aged sample and a young sample.

Correlation with other tests is another commonly used validation method whereby a new test should compare well with existing tests of the same trait. (The Binet test of intelligence – one of the earliest examples of this type of test – and the later Wechsler Adult Intelligence Scale were often used to validate new tests of intelligence.) Of course, any new test must have genuine advantages over what already exists. It might be easier to administer, more comprehensive, applicable to a broader sample; otherwise there is little point in developing something new.

Administration to extreme groups offers another method of validation, such that if two groups are known to differ markedly on a trait, the test should reflect this difference (a personality test, for instance, might clearly distinguish between previously identified extreme extraverts and extreme introverts). This is a particularly crude measure, however, since it will demonstrate only that a test is capable of identifying broad differences and not how well it measures fine distinctions (for instance, between mild and indeterminate introverts).

In fact, in the case of construct validity, a number of independent measures would be used to provide a comparison function, and most manuals for specific tests will offer extensive detail on how precisely the test was validated. See Box 6.8 for an example.

Box 6.3

A Closer Look At . . .

Test validity

An issue which is central to all psychological tests is the extent to which we can be confident that the test is measuring what it is supposed to be measuring. This is termed validity and there are many forms, as outlined in the main text of this chapter. Demonstrating validity, however, is a complex task, linked to the nature of the construct being measured – sometimes validity can be shown by correlating scores on a new test with scores on some existing or similar measure. A new arithmetical reasoning test in which the scope and depth of the concept is well understood can be matched to one of many existing instruments, or to scholastic assessments. With less-well-defined concepts, such as personality, intelligence or psychological wellbeing, it is more difficult: sometimes test assessments must be compared against interview evaluations, or longitudinal studies are required to determine the veracity of predictions made on the basis of a test, and sometimes research is required in several countries to demonstrate the cross-cultural validity of a measure.

In the manual of the General Health Questionnaire, or GHQ (Goldberg & Williams, 1988), a test designed to detect psychiatric disorders in community and non-psychiatric settings, a wide and comprehensive range of validation methods is described. In its development, the GHQ has been administered along with other measures of psychological health, such as the Profile of Mood States, or POMS (Worsley, Walters, & Wood, 1977), it has been matched to evaluations based on clinical interviews, and it has been used to predict GP consultations.

Table 6.5 is typical of the information available in the manual for the GHQ.

The GHQ is an extensively researched instrument which has been used in many situations and many cultures. For every application the manual provides details of validation research, with full citations and a commentary to aid the user in determining the efficacy of the instrument for current applications. Every student intending to use an instrument like the GHQ must become familiar with the relevant manual to determine the relevance of the instrument to particular groups in particular contexts.

Table 6.5 Correlation coefficients between scores on the GHQ-60 and a standard Clinical Interview Schedule (CIS) from three validation studies.

GHQ-60

Investigators	Year	Research interview	Correlation coefficient
Goldberg and Blackwell	1970	CIS	0.80
Goldberg	1972	CIS	0.77
Munoz et al.	1978	CIS	0.81

Adapted from the *Manual of the GHQ* (Goldberg & Williams, 1991, p. 44).

All of the foregoing discussion represents essential reading for anyone contemplating using a standard testing instrument as part of their study. For anyone who aims to make use of an existing test it is important to understand how it was devised, how reliable it is and what steps were taken to prove its validity. Familiarity with test norms is vital, since

this information tells us for whom the test is suitable, and what particular scores are likely to mean. Apart from being a major factor in determining your competence to use a given test, you will also be required in a final report to fully justify the use of a particular instrument. Moreover, it will be expected that, if a standardised test has been used in a study, you will be able to draw comparisons between your own findings and existing norms, and be able to discuss, knowledgeably, any deviations from published statistics. None of this is possible unless you *read the manual.* (We return to this issue in Part 7.)

While standardised tests will often comprise an element of undergraduate research, there will be occasions when no existing test is suitable for a particular design, or in which an issue is being explored using questionnaire or survey methods. This is usually the case when contemporary opinions, values and beliefs are being investigated. In such cases the researcher must develop her own instrument, a task sufficiently demanding that sometimes the development of a measure – with all the attendant requirements of reliability and validity – becomes the study itself.

Review

In this chapter we have considered a number of practical aspects of carrying out a study. By this stage you should now have a good idea of how many participants you require and how you will recruit them. You should also know precisely how you are going to collect your data – whether you will be using an existing measure or devising a measurement scale of your own. If you are developing your own instrument, you should appreciate the various options available in terms of measurement scales, the advantages of the different approaches and the associated pitfalls. You will also have sufficient familiarity with ethical guidelines governing psychological research to ensure that your study will be carried out in an ethical manner.

Suggested further reading

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct. American Psychologist, 57,* 1060–1073.

Asbury, J. E. (1995). Overview of focus group research. *Qualitative Health Research, 5,* 414–420.

This provides an informative overview of the history and use of focus group research.

Beech, J. R., & Harding, L. (Eds.). (1990). *Testing people: A practical guide to psychometrics.* Windsor, UK: NFER-Nelson.

A thorough but easy to read review of the major issues in psychological testing.

British Psychological Society. (2004). *Ethical guidelines: Guidelines for minimum standards of ethical approval in psychological research.* Leicester: British Psychological Society.

Provides a full account of the guidelines governing research with people.

British Psychological Society. (2005). *Code of conduct, ethical principles and guidelines.* Leicester: British Psychological Society.

In the event that some form of animal research is possible for undergraduates, this publication provides a complete set of guidelines governing the treatment and welfare of animals. This includes information on legal requirements.

de Vaus, D. A. (1996). *Surveys in social research* (4th ed.). London: University College London Press.

Chapter 15 on building scales provides a detailed review of the issues surrounding different kinds of measurement scales.