INTRODUCTION TO
Research Methods and Statistics in Psychology

A practical guide for the undergraduate researcher

SECOND EDITION

Ronald A. McQueen
Christina Knussen

ALWAYS LEARNING

PEARSON

A quick guide to . . . Which test should I use?

Selecting the most appropriate test for data analysis in quantitative research can be a daunting prospect for many. Below is a quick guide to choosing statistical tests which we hope will help, along with relevant chapter and section references. Each of the procedures given here can be viewed as a separate quick guide in the Appendix at the end of this book. Guides can also be downloaded and printed off as an additional resource from our Companion Website: **www.pearsoned.co.uk/mcqueen**

Ask yourself this question:
'Who is being measured or tested – different people or the same people being measured more than once?'

1. IF different people (e.g. different groups, samples, conditions)
   AND the outcome (dependent) measure is a category variable (e.g. Y/N response on a questionnaire item)
   THEN **crosstabulation** and **chi-square**. Part 4: Chapter 9; Section 9.12. Part 5: Chapter 11; Section 11.1.

2. IF different people (e.g. different groups, samples, conditions)
   AND the outcome measure is at least interval-scaled (e.g. ratings on a questionnaire item; scores on a test; a performance measure)
   AND there are only two groups or categories being compared
   THEN **independent *t*-test** (or **Mann-Whitney *U*–test** for non-parametric data). Part 5: Chapter 10; Section 10.11 (Section 10.12 for Mann–Whitney *U* test).

3. IF different people (e.g. different groups, samples, conditions)
   AND there are three or more groups or categories being compared,
   AND the outcome measure is at least interval-scaled (e.g. ratings on a questionnaire item; scores on a test; a performance measure)
   THEN **one-way ANOVA** (or **Kruskal–Wallis test** for non-parametric data). Part 5: Chapter 10; Section 10.13 (Section 10.16 for Kruskal–Wallis test).

4. IF different people (e.g. different groups, samples, conditions)
   AND the outcome measure is at least interval-scaled (e.g. ratings on a questionnaire item; scores on a test; a performance measure)
   AND there are two or more causal factors (e.g. Gender *and* Agegroup)
   THEN **two-way ANOVA**. Part 5: Chapter 12; Section 12.1.

5. IF the same people are being measured on more than one occasion
   AND there are only two repetitions
   THEN **paired sample *t*-test** (or **Wilcoxon's sign rank test** for non-parametric data). Part 5: Chapter 10; Section 10.17 (10.18 for the Wilcoxon test).

6. IF the same people are being measured on more than one occasion
   AND there are three or more repetitions
   THEN **one-way repeated measures ANOVA**. Part 5: Chapter 10; Section 10.19.

7. IF both independent and dependent variables are interval-scaled
   AND we wish to assess the nature and magnitude of the association between two continuous variables
   THEN **bivariate correlation**. Part 5: Chapter 11; Section 11.2.

8. IF both independent and dependent variables are interval-scaled
   AND we wish to predict one variable from the other
   THEN **simple regression**. Part 5: Chapter 11; Section 11.13.

9. IF both independent and dependent variables are interval-scaled
   AND we wish to predict one variable from several predictor variables
   THEN **multiple regression**. Part 5: Chapter 12; Section 12.4.

10. IF both independent and dependent variables are interval-scaled
    AND we wish to predict one variable from another, with the effects of other continuous variables controlled for
    THEN **partial correlation**. Part 5: Chapter 12; Section 12.13.

A closer look at . . .

Dichotomous scaling

Relying on a single item to provide data on an issue is clearly going to be of limited value – although many survey items are of this type when the aim of a question is simply to glean factual information (e.g. 'Do you possess a current driving licence?' Yes/No). When our interest is in something more complex, such as strength of feeling on an attitude issue, response to a single item is probably not going to be enough, especially as attitudes are quite complex and usually comprised of many elements.

Consider an attempt to assess the attitude of the student body towards their university course. We might conceivably ask only, 'Are you satisfied with your university course?' Yes/No. A single item like this, however, would fail to do justice to the many experiences which contribute to how someone feels about their life at university and it might be more useful to ask a series of questions . . .

Item # Do you find the content of lectures relevant to your course?	Yes/No
Item # Do you feel that seminars provide useful support for lecture material?	Yes/No
Item # Are academic staff approachable?	Yes/No
Item # Are online support materials easy to access?	Yes/No

. . . and so on.

In this way, the more Yes choices a person makes, the more satisfied they are. Indeed, it is possible with this format to incorporate sub-categories into the questionnaire, so we might have a few items on the delivery of course material, some items on the support systems available to students and perhaps even on the social facilities provided. Satisfaction scores for a given individual would comprise a simple count of the number of Yes responses, while scores for sub-populations would be expressed as mean scores (e.g. mean satisfaction scores for males vs females; arts vs science students; first year vs honours years, and so on). In the Eysenck personality scales, illustrated in the later sections on standardised tests, individual scores are based on this additive function, where a series of questions, such as, 'Are you a talkative person?' Yes/No, would combine to give an overall score on Extraversion, while a different series of questions would combine to provide an overall score on Neuroticism.

Multiple category scale items

Multiple category scale items offer three or more choices for the respondent.

Item # 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 7.3 demonstrates.

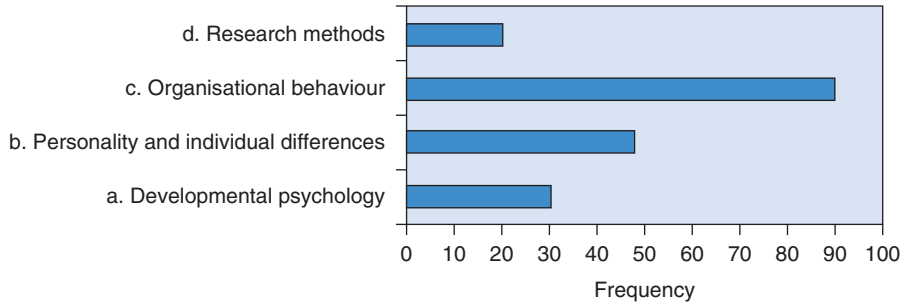

Figure 7.3 Frequency of response to the question, 'During the current semester which of the following modules have you enjoyed most?'

Rating scales

Rating scales 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 single category and require participants to indicate their strength of feeling on the issue. In the 'do you like the book?', example, participants could only give 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:

Item # I am enjoying this book:

Not at all Tremendously

1 2 3 4 5

Given our earlier discussion on leading questions, the astute reader might be anxiously raising their hand, trying to point out that by using the word, 'enjoying' we are assuming that enjoyment will feature somewhere in the reader's experience.

All right, we admit it, a leading suggestion if ever there was. Perhaps we should be more even handed, as:

Item # How do you feel about the book?

Hate it Love it

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, **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 7.2 and Figures 7.4 and 7.5 demonstrate this.

Table 7.2 Mean and standard deviation of response to an attitude question. Responses rated on a 5-point scale from 1 (not at all) to 5 (tremendously)

Item #	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. Figure 7.4 is one of a range of possible figures which might accompany these data. Here, 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

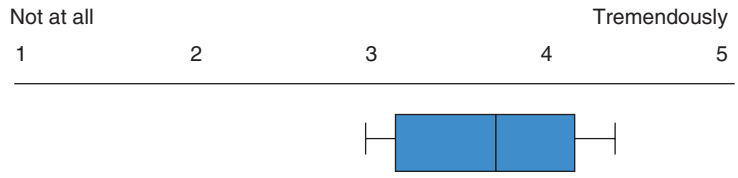

Figure 7.4 Boxplot illustrating responses on an attitudinal item

of the enclosed rectangle, is approximately 3.2 to 4.2; and the overall range of responses, shown by the 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. Figure 7.5 demonstrates this.

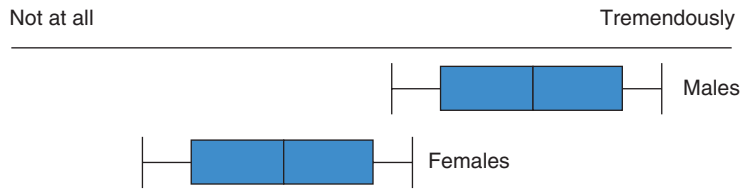

Figure 7.5 Responses of male and female participants to an attitudinal item ('I am enjoying the course')

From the literature

In the example from Loftus and Palmer (1974) at Section 7.3 above, participants were asked to recall whether broken glass was visible in the aftermath of a vehicle accident, depending on the particular verb used to describe the collision. They present their data as shown in Table 7.3.

Table 7.3 An excerpt from Loftus and Palmer (1974): Distribution of 'Yes' and 'No' responses to the question, 'Did you see any broken glass?'

	Verb condition		
Response	**Smashed**	**Hit**	**Control**
Yes	16	7	6
No	34	43	44

Source: Loftus and Palmer (1974), Table 2.

The findings are clearly presented in this table and indeed subsequent analysis showed that significantly more Yes responses were given when the verb 'smashed' was used to describe the incident than when a different verb was used.

Note that while the findings are self-evident from the table, there may be instances when a researcher will choose to highlight an element of the results in a figure. Based on the original data (Table 2 from the Loftus & Palmer 1974 study) the present authors offer two alternative methods of illustrating the leading-question effect – a simple bar chart (Figure 7.6a) and a line graph (Figure 7.6b). In both cases we have not shown the 'No' responses since our aim is to emphasise the finding that there were more misperceptions under the 'smashed' condition than any other.

→

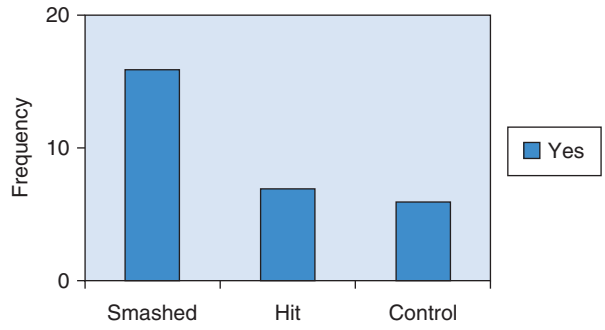

Figure 7.6a Chart showing frequency of misperceptions ('Yes' responses) under three verb conditions, based on the data in Table 7.3: bar chart

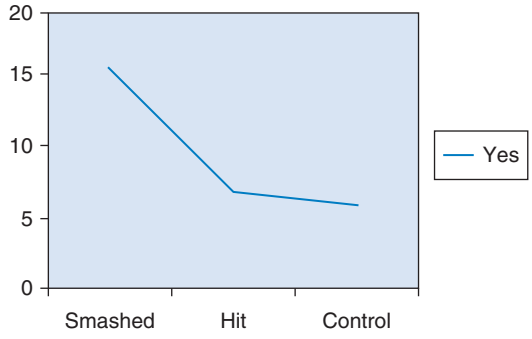

Figure 7.6b Chart showing frequency of misperceptions ('Yes' responses) under three verb conditions, based on the data in Table 7.3: line graph

Likert scales

A special variant of scaled items comes to us from Likert (1932). In the **Likert scale** 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, 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. De Vaus (2002) also provides useful background material on different types of scale. (Note that if these statistical terms are foreign to you, the next part of this book introduces basic statistical concepts.)

Likert scale, illustrated

This type of 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.

Item # Psychologists are nice people

Strongly agree	Agree	Undecided	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 qualitative 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. It is far better to treat the numbers as categories on an ordinal scale and avoid confusion altogether, which is what the purists would argue, and which was the original intent of this type of scale.

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 in the next section.