2 The experimental method

Thomas Gradgrind, sir. . . . A man of facts and calculations. A man who proceeds upon the principle that two and two are four, and nothing over, and who is not to be talked into allowing for anything over. Thomas Gradgrind, sir – peremptorily Thomas – Thomas Gradgrind. With a rule and a pair of scales, and the multiplication table always in his pocket, sir, ready to weigh and measure any parcel of human nature, and tell you exactly what it comes to.

(Charles Dickens, *Hard Times*)

One popular image of research is that it is concerned with formal experiments of various types (recall the statements of students at the beginning of Chapter 1). In this book we shall see that although experiments are important, they are by no means the only way in which research can or should be conducted. In this chapter, we shall consider what is meant by *the experimental method*. We shall also consider the use of *statistics* in research, and look at some of the more commonly employed statistical tools in applied linguistics. I should make clear at the outset that this chapter is not intended to teach the reader how to 'do' statistics. Rather, it is intended as a guide to the basic concepts needed to read with some understanding research reports utilizing statistics, and to appreciate the logic behind the use of statistical inference. For a more detailed introduction to statistics in applied linguistic research, see the references cited in the 'Further Reading' section at the end of this chapter.

This chapter addresses the following questions:

- What are *variables*, *samples*, and *populations*, and why are they important in research?
- What are the basic principles of sound experimental design?
- What do we mean by *inferential statistics?*
- When is it appropriate to use the following statistical procedures: *t-test analysis of variance, correlation, chi-square?*
- What is the difference between *true experiments*, *quasi-experiments*, and *pre-experiments?*

The context of experimentation

What are the contexts in which an experiment is the appropriate method for collecting and analysing data? Generally speaking, experiments are carried

TABLE 2.1 TYPES OF VARIABLES USED IN LANGUAGE RESEARCH

Type	*Example*
Nominal	L1 background: e.g., Arabic, Spanish, etc.
Ordinal	Rank on a test of grammar: e.g., first, second, third, etc.
Interval	Numerical score on standardised language test

out in order to explore the strength of relationships between variables. A variable, as the term itself suggests, is anything which does not remain constant. In our case, it includes language proficiency, aptitude, motivation, and so on. Language researchers often want to look at the relationship between a variable such as a teaching method and a second variable, such as test scores on a formal test of *language proficiency*. In such a case it is customary to distinguish between the two variables by giving them different labels. The label given to the variable that the experimenter expects to influence the other is called the *independent variable*. In our case this would be the teaching method. The variable upon which the independent variable is acting is called the *dependent variable* – in our case, the test scores.

Variables can also be classified according to the type of scale on which they are measured (see Table 2.1). A *nominal scale* measures mutually exclusive characteristics, such as sex and eye colour. (A subject cannot simultaneously belong to the category 'male' and the category 'female', or the category 'blue-eyed' and the category 'brown-eyed'.) *Ordinal scales* are for those variables which can be given a ranking, such as first, second, third, but in which the actual score itself is not given. An *interval scale* not only provides information on the rankings of scores, as does an ordinal scale, but also indicates the distance between the scores. Most test score data are of this type. A final type of scale for measuring variables is the *ratio scale*, which measures absolute values, such as temperature. Ratio scales are of little interest in applied linguistics, because variables such as language proficiency do not exist as absolute quantities; therefore, ratio scales will not be dealt with further here.

Let us consider an example of a situation in which an experiment might be an appropriate way of gathering data. Imagine that you have developed some innovative listening materials for low level learners. You have used these materials with a range of classes, and believe that they are significantly superior to the traditional materials which are used in your school. However, your colleagues are sceptical. How can you convince them that your materials are more effective than the traditional ones? There are many ways in which you could collect evidence. You could survey the students through interviews and questionnaires, and obtain their subjective impressions. You could ask a sympathetic colleague to become a participant observer in your classroom and make an ethnographic record of the teaching and learning

going on. These measures, however, are unlikely to sway your sceptical colleagues, who will be convinced only by test score data obtained through standardized tests.

You might be tempted to test your students at the end of the semester and present the results (assuming they are favourable) to your colleagues. However, you come across the following attack on such an approach (which is rather contemptuously dismissed as 'one-shot research'):

> Much research in education today conforms to a design in which a single group is studied only once, subsequent to some agent or treatment presumed to cause change. Such studies might be diagramed as follows:
>
> XO
>
> [X = the treatment administered to the subjects, and O = the observation.] [Unfortunately] . . . such studies have such a total absence of control as to be of almost no scientific value. . . . It seems well-nigh unethical . . . to allow, as theses or dissertations in education, case studies of this nature (i.e., involving a single group observed at one time only). (Campbell and Stanley 1963: 176–177)

Convinced by this argument, your next inclination is to test two groups, one which has used the innovative materials and one which has not. However, you quickly realize that it is no good simply testing the students at the end of the semester and comparing their scores with those obtained from another class at the same year level, because the groups might not have been at the same level to begin with. Fine, you might think, we can test both groups at the beginning of the term as well as the end. Then, if the group which has had the benefit of the innovative materials does better than the group that has used the traditional materials, we can presumably ascribe the superior performance to the materials.

While your research design is becoming more rigorous, it is still not rigorous enough to allow you to claim that there is a causal relationship between the independent variable (your innovative materials) and the dependent variable (the students' test scores). There is always the possibility that some factor other than the experimental materials has brought about the observed differences in the scores. For example, you may have happened to select a group of fast track or high aptitude students as the recipient of the experimental materials, and a group of slow learners as the 'traditional' group. In order to guard against such 'contamination', sound experimental design suggests that you should randomly assign students to either the control group, which uses the traditional materials, or the experimental group, which uses the innovative materials. You are then in a better position to argue that any differences on the end-of-term test are due to the experimental treatment (i.e., the innovative materials), because you can assume that other variables which might have an effect (such as intelligence or aptitude) exist in equal quantities in both the control and experimental groups, and therefore cancel one another out. You

should also test both groups of students before the experiment just to make sure that the groups really are the same.

If you carry out the procedures already described, that is, randomly assigning your subjects to either the control or experimental group, and administering a pre- and post-treatment test, then you could reasonably claim to have carried out what is known as a 'true' experiment. If you have not carried out these procedures, then the internal validity of your experiment is under threat (recall the discussion on reliability and validity in Chapter 1), because some variable you have not controlled may be affecting the dependent variable. Recalling van Lier's model (see Figure 1.3), you can see from this description why the formal experiment belongs to the highly controlled/highly selective quadrant of the diagram.

Unfortunately, it is not always practicable to rearrange students into different groups or classes at will. There are times when, if we are to carry out an experiment at all, it will have to be with intact groups of subjects, that is, subjects who have been grouped together for reasons other than the carrying out of an experiment. In these situations, while the internal validity of the experiment is weakened, it may still be thought desirable to proceed with the study. In instances such as this, researchers speak of quasi- or pre-experiments rather than true experiments. We shall look a littler further at these different types of experiments later in this chapter.

For argument's sake, let us imagine that you have been able to randomly assign sixty final-year secondary school students to control and experimental groups, and that a pre-test shows the two groups to be at the same level of proficiency. You teach both groups for a term, using the innovative materials with the experimental group and the traditional materials with the control group. At the end of the term, the groups are retested, and you obtain the scores for each student. You work out the *mean,* or average, for each group and obtain the following:

Control group: 58
Experimental group: 62

The experimental group has, on average, outscored the control group. Are you therefore in a position to claim that the innovative materials are superior to the traditional materials? Not yet. You have selected a sample, or subset, of all the possible students in the final year of secondary school who are studying your subject. If you tested them again tomorrow, or if you selected a second group of subjects and tested them, you would get different scores. Therefore, you need to use 'statistical inference' to work out whether the scores you obtained resulted from students' really being different, as suggested by the test scores, or whether the difference came about by chance or sampling variation. If all the students do share common, observable characteristics which differentiate them from other students, we say they represent a different *population.* A subset of individuals from a given population is a *sample.*

In order to illustrate the logic of inferential statistics, we need to go back a step or two and consider a number of basic concepts. This we shall do in the next section.

The logic of statistical inference

The aim of this section is to introduce you to the logic of statistical inference. While the information in the section will not necessarily provide you with the skills needed to carry out statistically based research, it should help you to understand the logic behind experimental research in which the researcher makes claims about an entire population based on data obtained from a subset or sample of that population.

In most research, it is not possible to collect data from the entire population of individuals in which one is interested. Consider an investigation of the listening proficiency of first-year secondary school French students. It would be extremely time consuming, although not impossible, to obtain data on all such students. Normally, someone wishing to carry out such an investigation would select a sample (say 30, 50, 100, 200) from the population and test these. However, a problem immediately arises: To what extent are the sample data representative of the population as a whole? Fortunately, certain procedures exist which enable us to determine the probability that the sample does represent the population from which it is drawn. In order to appreciate the logic behind these procedures, one must be familiar with the following statistical concepts: *mean, standard deviation, normal distribution,* and *standard error.*

From a statistical point of view, when studying numerical data of various sorts, the two things we will be most interested in are the extent to which the data are similar and the degree to which the data differ. The most frequently employed measure of similarity is the *mean* (symbolised by $\overline{X}$) which is simply the average of a set of scores (obtained by adding the individual scores together and dividing by the total number of scores). It gives us information about the central tendency of the scores. The *standard deviation* (SD), on the other hand, is the most important measure of dispersion, giving us information on the extent to which a set of scores varies in relation to the mean. It is calculated by deducting the mean from each individual score, squaring the resulting figures to get rid of the minus signs, adding these together, and dividing by the number of scores minus one. (Dividing by one less than the number of scores is a correction for the fact that the variability of scores for a single group of subjects tends to be less than the variability for all possible scores.) This gives us the *variance.* By obtaining the square root of this figure we arrive at the standard deviation. A simple example is set out in Table 2.2 to demonstrate the procedures involved.

From Table 2.2, we see that the variance for these scores is 5.111, and the