Joanna Bielska
University of Silesia

The experimental method in action research

1. Introduction

1.1. Experiment in action research — a paradigm clash?

Discussing the use of the experimental method in action research studies poses a considerable challenge. According to Dörnyei, action research, conducted by or in cooperation with teachers, has as its purpose 'gaining a better understanding of their educational environment and improving the effectiveness of their teaching' (2007: 191). Implemented by practitioners, it usually consists in systematically searching for a solution to some problem or puzzle encountered in the course of classroom practice. Therefore, action research, by definition, is said to be situational, that is concerned with the immediate context in which it is being carried out — the classroom. Consequently, generating knowledge that could be generalized beyond the classroom under investigation and would contribute to the theory of language learning is not a priority in the action research framework. Given that, and taking into account the variety and complexity of the factors influencing what goes on in the classroom, it is hardly surprising that in carrying out action research, teacher researchers draw extensively on qualitative research methods and data collection tools such as case studies, interviews, introspective accounts, diaries, journals, and observation. Although the data gathered in action research studies are sometimes presented in the numerical form, action research is far and foremost a grounded, process-oriented, qualitative endeavour.

The experimental method, in contrast, could be claimed by many to represent quantitative research 'at its most scientific' (Dörnyei 2007: 115). Actually, Brzeziński (2008: 9) claims that psychology has become an empirical science thanks to the experimental method, introduced as a means of testing hypotheses derived from theory. It is enough to have a brief look through his recent book on the use of the experimental method in the behavioural sciences (Brzeziński 2008) to realize how much methodological rigour is required in designing an experimental study and how closely the experimental method fits the quantitative research paradigm. As regards the SLA field, many of its original research methodologies were borrowed from the broader disciplines, especially psychology, sociology, anthropology, and linguistics, together with the principles underlying their reliable and valid adoption and refinement (cf. Chaudron 2005: 762). Consequently, in accordance with its position in psychology, the experimental method in SLA, viewed as part of the psychometric research tradition (Chaudron 1988, Brown 1988), has been consistently placed at the quantitative end of the qualitative-quantitative continuum of research methodologies (see e.g. Larsen-Freeman and Long 1991: 15). The experimental study, therefore, is primarily an ungrounded, outcome-oriented, confirmatory undertaking.

As can be seen from the above discussion, the incorporation of the experimental research methodology into action research arouses considerable controversy resulting from the incompatibility of their corresponding research paradigms. This controversy, however, can be minimised when the two approaches are perceived as complementary rather than contradictory and an experimental investigation is designed as part of a mixed methods study, where qualitative and quantitative methods of data collection and/or analysis are combined to increase the strengths and eliminate the weaknesses of its component methodologies. According to Mertens (2005, reported in Dörnyei 2007: 164), mixing methods has particular value when the issue under investigation is embedded in a complex educational or social context. Actually, any issue investigated within the action research framework can be said to be embedded in such a context. As noted by Nunan, 'classroom researchers appear to be increasingly reluctant to restrict themselves to a single data collection technique, or even a single research paradigm' (2005: 236—237). Greene and Caracelli (2003, reported in Dörnyei 2007: 168) are probably right in claiming that paradigm compatibility is of little relevance to actual empirical research, as inquiry decisions should be grounded primarily in the nature of the phenomena being investigated and the contexts in which the studies are conducted rather than in philosophical assumptions or beliefs. Therefore, the choice of the research methods, data collection

tools and analytical procedures should be centred around the research question. If the research question posed in an action research study requires verification of a research hypothesis through an experimental investigation, an appropriate type of experiment should be included in the research design. The data collected in the experimental investigation of language learning can be quantitative, qualitative, or both. The same concerns data analysis. Quantitative data will most often be subjected to statistical analysis, qualitative data — to interpretive analysis (cf. Grotjahn 1987, Nunan 1992). However, appropriate procedures of 'quantitizing' and 'qualitizing' data facilitate statistical analysis of qualitative data and interpretive analysis of quantitative data (see Dörnyei 2007 for the discussion of data transformation).

Taking all of the above into consideration, I would argue that viewing the experimental method from the perspective of mixed methods research, which has by now reached the status of 'the third research paradigm in educational research' (Johnson and Onwuegbuzie 2004, quoted in Dörnyei 2007: 167), gives researchers a wider repertoire of methodological options than considering it as a purely quantitative research method. Having said that, I must note that the qualitative aspects of the use of the experimental method will not be further discussed in this chapter, as the principles underlying qualitative data collection and analysis are common to other research methods and have been extensively discussed in several other chapters included in this volume.

1.2. Experiments in classroom research — contributions to theory and practice

The use of the experimental method in the study of second language acqusition has a long and rich tradition. Experimental and quasi-experimental studies have been successfully conducted both inside and outside language classrooms and generated a large body of reliable findings. A good example of an issue that has been investigated with the use of the experimental methodology is the efficacy of implicit feedback in the form of so-called corrective recasts which has been researched through true experiments conducted in laboratory settings (see e.g., Ortega and Long 1997, Mackey 1999) as well as quasi-experiments conducted in classroom settings (e.g., Doughty and Varela 1998, see Long 2007 for an extensive review of this work). Although the data collected in such studies are frequently gathered in the language classroom and the research findings might in the long run have some effect on designing language teaching, the

value of such research rests primarily on its contribution to the theory of second language acquisition rather than on the direct teaching implications it offers to classroom teachers.

In the case of experiments carried out within the action research framework, however, the above-mentioned hierarchy of goals is reversed. The value of research findings produced in action research studies is judged primarily on the basis of their usefulness for the classroom, be that the enhancement of a teacher's practice (Allwright and Bailey 1991, Allwright 1993, Wallace 1998) or the more collaboratively effected larger-scale methodological change (Crookes 1993, Van Lier 1996, Burns 1999, 2005, Roberts 1998). If the methodological principles of research design are followed and the generalization of findings beyond the immediate context is feasible, the study results may contribute to SLA theory building; such contribution, however, is not a defining feature of action research. As noted by Long, 'SLA theories might provide insight into putative universal *methodological principles* for language teaching [...] while saying little or nothing about the inevitable and desirable particularity of appropriate classroom *pedagogical procedures*, in which the local practitioner, not the SLA theorist, should always be the expert' (2007: 19). Action research, therefore, can be viewed as a means of gaining the necessary expertise by 'the local practitioner', i.e. a language teacher. To use the above mentioned example of corrective recasts again, Long (2007) notes that while a SLA theory might hold the provision of negative feedback to be necessary or facilitative in the process of second language acquisition, the choice of appropriate pedagogical procedures, ranging from the most implicit corrective recasts to the most explicit forms of error correction, rests entirely on the teacher working in a particular educational context. Experimental studies in action research, therefore, serve the purpose of enabling teachers to make, evaluate, or justify their choices concerning classroom instruction by testing hypotheses related to the contextualised use of pedagogical procedures. Many experimental studies conducted within the action research framework are designed to test some methodological innovation, be that a set of more innovative teaching techniques, a selection of some modern teaching materials, a new coursebook, software package or course module. As noted by Komorowska (1982: 111), the purpose of such an experiment is the optimalization of the teaching and learning process rather than contribution to the theory of language acquisition. Viewed from this perspective, the classroom is better considered not as an 'experimental laboratory' for SLA research, but as a 'social context for language learning' (Breen 1985, reported in Burns 2005: 245). In order to provide valid, meaningful results, experiments should be conducted within such educational contexts rather than in artificially created settings.

The two perspectives on the use of the experimental method in language learning and teaching have some consequences regarding the methodology of conducting experimental research. Psycholinguistic experimental research in SLA, most often designed to assess the impact of an isolated variable on the development of L2 competence, is characterized by the methodological rigour required of the studies whose major objective is testing the scientific hypotheses derived from theory. Experiments conducted within the action research framework, on the other hand, being educational rather than psycholinguistic in character, are often characterised by a certain degree of methodological flexibility. For example, rather than isolating a single variable to be studied with respect to its effect on the learners' interlanguage development, an experimental study in education may involve 'assessing the effect of a whole range of classroom activities which together combine a cluster of theoretically motivated characteristics, and which are put together in an educationally viable way' (Harley 1989: 331). Exactly to what extent the conservative principles of the experimental method can be compromised in educational research without seriously threatening its validity is a matter of considerable controversy. Some of the arguments raised in this debate will be discussed in the remaining parts of this chapter.

2. Description of the method

2.1. What is an experiment?

Every day all of us make observations that can easily be described, but that are not necessarily equally easy to explain. In the process of language teaching, a teacher might note, for example, that some learners make more spelling mistakes than others, that in spite of all the speaking activities done in the classroom learners' fluency does not seem to improve, or that one group of learners clearly outperform another group on achievement tests although they are taught by the same teacher, follow the same syllabus, and have the same number of hours of English per week. While describing or even documenting such observations is relatively straightforward, establishing the causes for the phenomena in question may pose considerable difficulty. For instance, slow progress in the development of the learners' discourse competence may be caused by insufficient exposure to language used in authentic situations, low motivation or language aptitude of the

learners, too little practice, language anxiety and many other factors; it may also be caused by several interacting factors. If the teacher decides to search for the causes in a systematic way by obtaining, analysing and interpreting data related to the problem, the 'Why?' question becomes a *research question*. Usually, we ask questions when we do not know the answers. Actually, as noted by Hatch and Lazaraton (1991: 23), that does not mean that we have no idea about what those answers might be. After all, in the above example we have been able to enumerate a few potential reasons for the slow development of the learners' discourse competence, based on the theoretical knowledge of the topic and classroom experience. We could, therefore, try to formulate a statement that would be a potential answer to the research question. In formal terms, such a statement about an expected outcome of the research is called a *research hypothesis*. The aim of the research process is collecting the evidence that will support or not support the hypothesis (or hypotheses) stated in a research study. Such *hypothesis testing*, however, would not be possible without carefully controlling the variables under study, especially when unambiguous cause and effect relationships are to be established. As Dörnyei points out, 'to establish firm cause-effect relationships is surprisingly difficult because in real life nothing happens in isolation and it is hard to disentangle the interferences of various related factors' (2007: 115). The experimental method offers a way of such disentangling by introducing research designs in which the target variables are measured and manipulated while all the other variables that might influence the relationship under study are carefully controlled. An experiment, therefore, can be defined as a hypothesis-testing procedure designed to establish, in a controlled environment, the existence and strength of cause and effect relationships between variables.

As can be seen from the above definition, the notion of a 'variable' is central to the discussion of experimental designs. Although properly identifying, defining and classifying variables constitutes an essential element of any good research study, the experimental method is particularly sensitive to the lack of precision in this respect. Using van Lier's (1988) terms, a formal experiment is characterized by a high degree of intervention and a high degree of selectivity and thus falls into the 'controlling' space in his model of research designs (see also Nunan 1992: 5—7). As such, it requires that all the variables to be focused on should be specified prior to the study. Moreover, one of the essential characteristics of a properly designed experimental study is its replicability (Brown 1988). A study can be described as replicable if it can be reproduced under similar conditions by an independent researcher, which is only possible if a clear, explicit, and complete report of the original study is available. Such a report should contain precise definitions of all the variables under study

together with the description of the proposed relationships between them. The next section, therefore, deals with the ways of defining, measuring and classifying variables.

2.2. Describing variables

2.2.1. Definition and operationalization

A *variable* can be defined as 'something that may vary, or differ' (Brown 1988: 7). More specifically, a variable is an attribute of a person or an object that can take on different values, i.e. 'vary', from person to person, text to text, object to object, and/or from time to time (cf. Howell 1999, Hatch and Lazaraton 1991). Some human characteristics such as gender or first language background differ from person to person but do not generally vary over time, other characteristics such as age, motivation, or language proficiency vary both over time and between individuals. While some of the differences between people can easily be observed, others are not directly observable. Dividing a group of people into two groups on the basis of their gender or ordering a group of students according to their height can be done without much difficulty. But what if you were to rank them on the basis of their motivation, anxiety, or intelligence? You would have to stop and think what it really means to be motivated, anxious or intelligent. It would soon turn out that the task poses considerable difficulty. As noted by Brown (1988), it is important to distinguish variables, i.e. what can be observed of the human characteristics, from the underlying constructs that they represent. For example, a learner will be considered gifted when he or she shows certain characteristic behaviours or abilities associated with giftedness such as solving difficult analytical tasks in a very short time, composing a symphony at the age of five, or acquiring a foreign language at a very high rate. While giftedness or language aptitude are broad abstractions, composing a symphony or successfully working out the rules of an artificial language are observable signs of the actual human characteristics.

Designing a good experimental research study requires that the theoretical terms used to describe abstract characteristics can be translated into the language of empirical terms. The process of searching for the empirical equivalents of theoretical terms is called the *operationalization* of variables (cf. Francuz and Mackiewicz 2007: 54). In Brown's words, 'an operational definition should take a variable out of the realm of theory and plant it squarely in concrete reality' (1988: 8). This is a critical stage in the process of constructing any experimental study. The operational definitions

used in the study will have direct influence on the internal and external *validity* of the study. If they are adequate descriptions of the constructs in question, that is, if 'pieces of data collected to represent a particular construct really succeed in capturing the construct' (Hatch and Lazaraton 1991: 38), the validity of the study will be increased. In other words, you will feel more confident in any claims you decide to make.

Francuz and Mackiewicz (2007: 56) note that we can talk not only about theoretical and empirical terms, but also about theoretical and empirical sentences. The contruct validity of the operational definitions used in the study will have a profound effect on the relationship between the two types of sentences. Research questions as well as the answers to them are usually formulated as theoretical sentences, i.e. ones containing at least one theoretical term. In contrast, the results of an experiment take the form of empirical sentences, i.e. ones that do not contain any theoretical terms. Inappropriate operationalization, therefore, may result in the conclusions being questioned. For instance, if you found in a study that the group of subjects whose errors were consistently corrected obtained higher scores on a multiple-choice vocabulary posttest than the group of subjects whose errors were ignored, could you safely conclude that error correction facilitates vocabulary acquisition? If you did, your critics might argue that vocabulary acquisition is something more than performance on a written vocabulary test composed of multiple-choice items and claim that your conclusions are unjustified. As many contructs relevant to language learning and teaching are extremely difficult to define in empirical terms, the operational definitions should always be kept in mind when you read, evaluate, and compare the results of research reports (Hatch and Lazaraton 1991). Designing and conducting an experimental study is a laborious and time-consuming process. However, no matter how much time and effort is invested in data collection and analysis, the conclusions will always be limited by the operational definitions used in the study. As stressed by Francuz and Mackiewicz, 'operationalization is the key to the experiment' (2007: 54).

2.2.2. Measurement

Experimental studies are designed to test hypotheses concerning cause-effect relationships between variables. Assessing the strength of such relationships with any degree of confidence would not be possible without appropriate measurement of the variables under study. As in other types of quantitative research, data collection procedures used in experimental studies yield numerical data which are then subjected to different types of statistical analysis, including hypothesis-testing procedures. All

the variables included in experimental designs, therefore, need to be quantified. However, many variables that are of interest to educational psychologists and applied linguists cannot easily be observed, let alone precisely measured. The choice of the appropriate method of measurement is not always straightforward. As noted by Hatch and Lazaraton, 'the way you measure variables will depend in part on the variable itself and its role in the research, and in part on the options available for the precision of measurement' (1991: 59). The description of the way in which a variable is to be measured in a given study should always be included in its operational definition.

Depending on the method of measurement, variables can be divided into two major classes: *continuous* variables (e.g. test scores) in which the variable can take on any value between the lowest and highest points on the scale, and *noncontinuous* (discrete) variables (e.g. gender, type of school) in which the variable can take on only a relatively few possible values (cf. Howell 1999: 20). This division, however, is very general and does not take into account some essential differences in types of measurement. A very useful and widely accepted division of variables based on the types of measurement was proposed by an American psychologist S.S. Stevens (1951), who distinguished four types of scales (nominal, ordinal, interval, and ratio), also called levels of measurement, as they can be arranged hierarchically according to the degree of precision in measurement (cf. Brown 1988, Howell 1999, Brzeziński 2008). Data from more precise measurement can be converted to data from less precise measurement, but not vice versa.

Nominal scales are generally used for the purpose of classifying categorical data, that is naming categories and assigning category labels to observations. Nominal scales can be dichotomous (when the variable has two levels, e.g. yes/no) or consist of more than two categorical values. For example, the nominal variable 'type of school' might have three levels (1 = primary, 2 = lower secondary, 3 = secondary), the variable 'proficiency level' might have two, three, four or more levels depending on the framework of reference. The categorical values on nominal scales are qualitative in character. The classification numbers often used to code data are labels representing levels of the nominal variable and have no arithmetic value. The observations falling into the nominal scale categories can be tallied, resulting in so-called *frequency* counts, which show how often something occurs in the data. These frequencies do have arithmetic value (for a more detailed discussion see Hatch and Lazaraton 1991).

Ordinal scales are used to order people, objects, or events along some continuum (Howell 1999: 16). For example, individual students (or groups of students) can be ranked from best to worst on the basis of an achievement

test, or from 'very unmotivated' to 'very motivated' on the basis of self-report data. The defining feature of ordinal measurement is that it 'orders responses in relation to each other to show strength or rank' (Hatch and Lazaraton 1991: 57). While the numbers used in an ordinal scale have arithmetic value, no information is given about the intervals between the points on the scale.

Interval scales are scales with equal intervals between scale points. Apart from ordering data, they provide information as to the actual distance between the ranks. The general assumption is that each interval means an equal increment. Typical data generated with the use of interval scales are test scores, especially on standardized tests. It is a matter of some debate whether the scales typically used in questionnaires such as a 4-, 5-, 7- or 9-point scales ranging from 'never' to 'always', 'very unhappy' to 'very happy' or, as in Likert-type scales, 'strongly disagree' to 'strongly agree', should be considered interval or ordinal scales (see e.g. Francuz and Mackiewicz 2007, Hatch and Lazaraton 1991). The issue is important as it has an effect on decisions concerning statistical analysis of the data. The general tendency is to use wider scales as they more closely approximate equal intervals.

Ratio scales have all the properties of interval scales, but they differ from them in two aspects: they have an absolute zero value (while the lowest value on an interval scale is arbitrary) and, as noted by Brown (1988: 22), the points on the scale are precise multiples of other points on the scale. While 20 seconds is twice as long as 10 seconds, and 0 seconds means no time at all, we cannot really say that a person with a test score of 40 points knows twice as much English as a person who scored 20 points, and zero points on a test do not indicate total lack of language competence. For these reasons, ratio scales are not usually used in applied linguistics research.

The choice of the method of measument for the variables to be studied in a research project is an important decision on the part of the researcher, as it will influence the ways in which data are arranged, described and analysed. Consequently, it will have an effect on the final conclusions that can be drawn from the study.

2.2.3. Function

Properly designing an experimental research study involves not only identifying all the variables that relate to the research question, but also classifying them with respect to their functions. Experimental studies are designed to investigate how some variables *affect* other variables, so decisions concerning their roles in a research design, based on the

researcher's intuitions as to the links between the phenomena in question, will directly influence the logic of the study.

The basic functional division of all the variables in a study involves classifying them as either dependent or independent. The *dependent* variable is the central variable in a study. It is believed to 'depend' on other variables, called *independent* variables, in the sense that the independent variables affect or cause a change in the dependent variable. While the dependent variable is always, in Brown's words, 'the variable of focus' (1988: 10), the independent variables may be further subdivided according to the nature of their relationship with the dependent variable and the researcher's intentions concerning their place in the study design. The basic link to be investigated in the study is that between the major independent variables and the dependent variable. In an experimental study, the major independent variable will be manipulated in order to determine (and measure) its effect on the dependent variable. Some independent variables will influence the dependent variable indirectly, by mediating or moderating the relationship between the independent and dependent variables. If such a variable is included in the study design, it is called a *moderator* variable. If, however, it is not included in the study because it has not been or cannot be identified, it is called an *intervening* variable (cf. Hatch and Lazaraton 1991, see also Brown 1988 for a different definition of an intervening variable). The last type of independent variables are *control* variables, that is variables which might affect the dependent variable, but are not of central concern in a particular research project, so their potential effect on the outcome needs to be controlled (Hatch and Lazaraton 1991). In the case of nominal variables, this can be done by limiting the study to just one level of the control variable, at the same time, however, limiting the generalizability of the study. In the case of continuous variables, their effect can be controlled with the use of appropriate statistical procedures at the stage of data analysis.

Not all variables influencing the dependent variable can be identified and taken into account in planning a research design. As stressed by Hatch and Lazaraton, 'in all research, we can only account for some portion of the variability that we see in the major, dependent variable' (1991: 68). There will always be factors that have not been considered and might be the cause of 'error' in the study. However, care should be taken to minimize that error by carefully analysing the context of the study so that the most important sources of variability in the dependent variable are identified and taken into account in planning a research project.

2.3. Types of experiments

The major aim of conducting experimental research in the behavioural sciences is to predict and explain human behaviour. To that end, as has been noted before, experiments are designed with the purpose of establishing causal relationships between variables. In the case of classroom research that usually means identifying the causal link between some treatment (e.g. a new teaching strategy) and its consequence (e.g. language learning outcomes). According to Dörnyei, 'the main strength of the experimental design is that it is the best method — and some would claim the *only* compelling method — of establishing cause-effect relationships and evaluating educational innovations' (2007: 120). In order to establish such links in a valid and generalizable manner the experimental design has to satisfy certain criteria.

The first one concerns *random selection* and *random assignment* of subjects. The principle of random selection means that every member of the target *population*, i.e. all the people to whom the results of the study will be generalized, should have an equal chance of being included in the *sample*, i.e. the group of people who will actually be studied in the experiment. Random assignment, on the other hand, concerns assigning subjects to the comparison groups included in the experimental design in that each member of the study sample should have an equal chance of being included in any of the experimental or control groups used in the study. The main goal of random selection is assuring the representativeness of the sample with respect to the population, so that generalization of the research findings is justified. In other words, random selection enhances the *external validity* of the study. The role of random assignment is to eliminate any preexisting differences between the comparison groups in order to assure their equivalence, so that any effects found in the study can be attributed to the independent variable. Random assignment will therefore enhance the *internal validity* of the study and, indirectly, its external validity (a study which does not have internal validity cannot have external validity). The second criterion also deals with strengthening the internal validity of the study as it requires establishing the baseline for comparison. Apart from the *experimental group* subjected to the treatment under investigation, at least one *control group* receiving the unmarked (or standard) treatment should be included in an experimental study.

The experimental design that satisfies both criteria is called a *true experimental* design. It has experimental and control groups and random selection and assignment of subjects. Unfortunately, in educational contexts, random selection and random assignment of subjects are hardly ever possible. Moreover, as noted by Komorowska (1982: 109), limiting

the use of the experimental method in language teaching methodology to experiments conducted with random selection at the level of individual subjects rather than classes or schools is actually disadvantageous. All in all, true experimental designs are rarely used in applied linguistics research, let alone action research.

Most of the experimental studies in our field use *quasi-experimental* designs, which have experimental and control groups but no random assignment of subjects. These studies are usually conducted with so-called *intact* groups, formed for purposes other than the research project. Since the randomization procedures have not been observed, the validity of the study is under threat. Researchers differ as to whether quasi-experimental designs can establish causal links and whether their findings can be generalized. Hatch and Lazaraton, for example, claim that 'when random selection is not possible, causal claims are also impossible' (1991: 85) and that 'we cannot *generalize* anything from the results unless we have appropriate subject selection' (1991: 42). In their view, quasi-experimental designs can only be used to give evidence in support of links between variables for the classes participating in the research. In contrast, other researchers (Johnson and Christensen 2004, Dörnyei 2007) believe that randomization is not the only way of eliminating threats to validity, and that if certain procedures are applied to assure the representativeness of the sample and the equivalence of the control and experimental groups, a quasi-experimental design can produce meaningful findings which can be generalized beyond the immediate context of the study. As Dörnyei concludes, 'as a result, it is generally accepted that properly designed and executed quasi-experimental studies yield scientifically credible results' (2007: 118). However, a detailed description of the population to which the findings will be generalized must be provided in order to avoid the risk of overgeneralization.

The last type of experimental research design, which fails to meet both criteria, is a *pre-experimental* design. In pre-experiments there is no control group, and, consequently, no random assignment of subjects. As a result, pre-experimental designs are incapable of generating data necessary to test research hypotheses. They can, however, be quite useful in piloting materials or testing procedures to be used in a more complex research study. They also provide useful insights and generate hypotheses concerning language learning and teaching, which can later be tested with more rigorous research methods.

2.4. Choosing a research design

The number of options available in selecting an experimental research design is quite impressive. In their classic text on experimental research, Campbell and Stanley (1963) list sixteen different design types. However, the experimental designs traditionally used in educational research are more limited in number, the most commonly used ones are briefly discussed below. What has to be stressed, though, is that there is space for creativity in designing an experimental study. In planning a research project, you should first and foremost consider your research question and then choose or construct a design which will help you generate valid and reliable findings. In the schematic representations used below, *X* stands for the independent variable (treatment), *Y* for the dependent variable, *E* for the experimental group, *C* for the control group.

2.4.1. One-shot design

X — Posttest *Y*

The so-called one-shot design is a type of pre-experimental design. This type of experimental research involves giving treatment to one group of subjects followed by a test (note that the word *test* is a cover term for any type of measurement). The general idea is that the effectiveness of the treatment can be evaluated by the test. Unfortunately, even if the treatment and measurement are carefully documented, the findings may not be valid. The results obtained on the test may be the effect of the treatment or may be attributed to factors other than the treatment, such as the subjects' history or maturation. The only thing that Ms Brown (see Figure 1) can do is describe the data and reflect on it. However, even if it seems that the new handbook has worked, continuing with its use or recommending it to other teachers could not be justified by the data.

Disappointed with the coursebook she had been using for several years, Ms Brown, an EFL teacher in a lower secondary school, decided to try out a new, more innovative coursebook with the class of first-graders she was about to start teaching. She chose the book and followed it closely for two semesters. In order to evaluate the effectiveness of the coursebook, at the end of the school year she administered a test to see if her students met the objectives set for the course.

Fig. 1. An example of one-shot design

2.4.2. One-group pretest-posttest design

Pretest *Y* — *X* — Posttest *Y*

The one-group pretest-posttest design constitutes an improvement over the one-shot design in the sense that the researcher can rest assured that the subjects did not know the material tested on the posttest prior to the treatment. On the other hand, however, administering the pretest may threaten the validity of the study due to the so-called *practice effect*, the subjects may simply learn the material included in the pretest, which will influence the scores they get on the posttest. Moreover, the pretest may give the subjects some idea as to what they are supposed to learn during the study. As a result, the study outcomes might be influenced by *subject expectancy*. If Ms Brown (see Figure 1) included a pretest in her study, she could make sure that the change between pretest and posttest scores occurred due to the factors working within the duration of the study rather than prior to it. The claim that the use of the handbook was the only factor causing that change would not, however, be justified. The one-group pretest-posttest design, therefore, cannot be used to test research hypotheses concerning relationships between variables. It can, however, be used to gather preliminary data through pilot studies. As in any other design where the pretest is used, the researcher should control for practice effect and subject expectancy.

2.4.3. Control group pretest-posttest design

Pretest *Y (C)* — X_1 — Posttest *Y (C)*

Pretest *Y (E)* — X_2 — Posttest *Y (E)*

The control group pretest-posttest design is probably the most popular experimental design. Through including a control group, the researcher eliminates some of the threats to the internal validity of the study by introducing the baseline for comparison. Underlying the logic of this design is, in Dörnyei's words, 'a simple but ingenious methodological idea' which he describes as follows:

> First, take a group of learners and do something special with/to them, while measuring their progress. Then compare their results with data obtained from another group that is similar in every respect to the first group except for the fact that it did not receive the special treatment. If there is any discrepancy in the results of the two groups, these can be

attributed to the only difference between them, the treatment variable (Dörnyei 2007: 116).

An example of the above procedure is given in Figure 2. The strength of this experimental design comes from the fact that it involves both within-groups and between-groups comparisons. What can be established is not only whether the experimental treatment worked, but also whether it actually worked better (or worse) and how much better (or worse) than the standard treatment administered to the control group.

Having taught two groups of third-grade grammar school students for eight weeks, Ms Bird, an EFL teacher, administered a proficiency test to check her students' progress towards proficiency level B1, which they were required to reach by the end of the semester. Analysing the test results, she noticed that in both the groups the students' scores on the listening comprehension component of the test were significantly lower than those on the remaining parts of the test (Reading, Use of English, Writing). She had a feeling that although the coursebook she had been using contained quite a lot of listening comprehension tasks, they were very artificial and relatively simple, probably not contributing much to the development of the students' listening skills. Before asking the students to buy a new book, however, she wanted to make sure that her interpretation of the reasons for the low listening scores was correct. In order to check whether it was really the listening comprehension material included in the book that could be blamed for the disappointing results and at the same time in an attempt to remedy the situation by introducing a specially prepared set of more challenging listening comprehension tasks based on the recordings of authentic language, Ms Bird designed a quasi-experimental study to be conducted over the next eight weeks. The two groups she had been teaching were randomly designated as the experimental group and the control group. For the next eight weeks the control group received regular instruction based on the coursebook, whereas in the experimental group the listening comprehension tasks included in the coursebook were replaced by the new tasks prepared by Ms Bird. At the end of the treatment, another B1 proficiency test was administered, equivalent in form and content to the test administered prior to the study. The results of the study were then subjected to statistical analysis, which confirmed Ms Bird's intuitions.

Fig. 2. An example of control group pretest-posttest design

Ingenious as it may be, the classic experimental design is not without problems. As has been mentioned before, assuring the equivalence of the control and experimental groups is not always possible. The potential effect of the pretest on the posttest scores needs to be taken into account as well. Moreover, the results may be influenced by the so-called *Hawthorne effect*. The experimental treatment in educational research usually involves some innovation, which is likely to be more attractive to the subjects than regular instruction. The differences in performance between the control and experimental groups may be, at least to some extent, caused by the pleasure of being treated in a special way rather than by the character of the treatment itself. Moreover, the different outcomes may be caused by *researcher expectancy*. The researcher may be so focused on trying to 'prove' that the innovation is superior to the

standard that it may influence the way the treatments are administered, resulting in biased outcomes. Care should be taken to minimize the effect of such extraneous variables (see Brown 1988 for a useful discussion of the ways to control extraneous variables).

Despite the above limitations, the control group pretest-posttest design is a powerful design, capable of providing evidence in support of cause-effect relationships. The classic experimental design has many variants. It may include more experimental and/or control groups. It can also be improved by introducing delayed posttests in order to assess the long term effects of the treatment.

2.4.4. Time-series design

Pretest 1 *(Y)* — Pretest 2 *(Y)* — Pretest 3 *(Y)* — X_1 — Posttest 1 *(Y)* — — Posttest 2 *(Y)* — Posttest 3 *(Y)*

Time-series designs are used when for some reason including a control group is impractical or impossible. Since that is often the case in action research, where quite often the data are collected from one intact group, time-series designs are very useful in research conducted by practitioners. In time-series designs, the class is its own control group. First, in order to establish the normal growth in performance over a period of time, several pretests are taken by the subjects. As the next stage the experimental treatment is administered, followed by a series of posttests. A sudden growth (or drop) in performance following the treatment would indicate its effect. There are many variations in time-series designs (see e.g. Hatch and Lazaraton 1991, Komorowska 1982). Moreover, they can be combined with designs using control groups, which would maximize the advantages and minimize the drawbacks of both types of designs.

The major advantage in time-series designs is that they document the learning process rather then focusing only on its final product. Additionally, as the same group serves both as the control and the experimental group, the threat of using non-equivalent groups is eliminated, which enhances the validity of the study. There are, however, a few disadvantages of the time-series design. For example, multiple testing increases the risk of the measures influencing each other (practice effect) or, especially when the tests measure attitudes rather than learning outcomes, causing a change in the subjects (reactivity effect). Moreover, time-series designs are longitudinal designs, so the data collection process requires more time than an equivalent design with a control group. An example of a study based on the time-series design is presented in Figure 3.

Mr Edwards noticed that the compositions of his third-grade lower secondary school students were full of errors in the use of tenses, although the students did not usually make them when filling in grammar exercises in their workbooks. He had an idea that a series of workshops integrating grammar instruction with reading and writing activities would remedy the situation. In order to evaluate the effectiveness of the workshops, Mr Edwards decided to conduct an action research study using a time-series design. Every two weeks he asked his students to write an in-class composition in which they were to narrate a story. Analysing the data, Mr Edwards calculated the error rate for each student (the number of errors in tensed verbs was divided by the total number of finite clauses used in the composition). After six weeks the students participated in three 90-minute workshops that he had prepared. In the six weeks following the workshops the students wrote another three compositions, one every two weeks. The comparison of mean error rates on all six compositions revealed a rapid drop following the workshops. Mr Edwards was pleased to see that the workshops had a positive effect on his students' outcomes.

Fig. 3. An example of a time-series design

2.5. Analysing data

A detailed discussion of quantitative data analysis falls beyond the scope of this chapter. The choice of statistical tests will depend on the research question, the type of experimental design and the type of data collected in the study. Regardless of the research design, however, the statistical procedures used in quantitative research can be divided into two different types defined by their function. *Descriptive statistics*, as the name suggests, are used to 'describe' a set of data, to organize and summarize it in a logical way so that it becomes interpretable. *Inferential statistics*, on the other hand, are used to give us confidence in any general claims we want to make on the basis of what has been observed in the sample. In other words, inferential statistics are used to guard researchers against making unreasonable generalizations on the basis of limited data. While descriptive statistics will be used to analyse data in any quantitative study, the use of inferential statistics is allowed only when the sample can be regarded as representative of the population to which the findings are to be generalized.

2.5.1. Descriptive statistics

There are three types of descriptive statistics: frequency counts, measures of central tendency, and measures of variability. They will be discussed in turn.

Frequency counts

Frequency data show *how often* a variable is present in the data. The data are noncontinuous and describe nominal (discrete) variables (Hatch and Lazaraton 1991: 62). The simplest way of describing a set of data that are frequencies are frequency counts, which can be summarized in tables or presented graphically in the form of bar graphs, histograms, frequency polygons, etc. In this way you can display, for example, how many subjects in your sample fell in a particular category (e.g. male/female, beginner/intermediate/advanced, etc.). When you have a relatively small data set, the raw frequency totals can usually be used. With larger data sets, however, reporting relative frequency such as percentage or proportion may be more informative (for more details see Hatch and Lazaraton 1991, Howell 1999).

Measures of central tendency

Score data show *how much* of a variable is present in the data. The data are continuous and describe ordinal or interval variables (Hatch and Lazaraton 1991: 62). Plotting the values of the dependent variable against their frequency of occurrence results in the *distribution* of values in the data. Measures of *central tendency*, the mean, the median, and the mode, are used to describe the central point of the distribution.

The most commonly used measure of central tendency is the *mean*, which is the sum of the scores divided by the number of scores in a data set, that is, the arithmetic average of the scores in a data. The mean takes all scores into account and is the only measure of central tendency which can be manipulated algebraically. It may, however, be seriously affected by extreme scores, so before calculating the mean it is always advisable to examine the distribution for *outliers* (scores that stand out from the rest of the distribution).

The *median* is the score which is the midpoint of the distribution. In other words, it is the middle value in the scores arranged in numerical order. Half of the scores fall below the median, and half above it. If the number of scores is odd, the median is the middle score, if the number of scores is even, the median is the average of the two middle scores. The median is not seriously affected by extreme scores. It is often used when the data describe an ordinal variable.

The *mode* is the most commonly occurring score in a set of scores. It is most easily obtained by drawing a frequency polygon of the data and finding the score corresponding to the peak of the polygon. As noted by Hatch and Lazaraton (1991: 161), the mode is the measure of central tendency which is most seriously limited because it is easily affected by chance scores. What is more, the mode depends on how the data are grouped. However,

exploring the distibution for the number of 'peaks' is essential, since it may be, for example, bimodal or trimodal rather than unimodal, a feature which the mean or the median will not detect.

As an example, consider the following data set: 3, 5, 7, 9, 11, 9, 5, 10, 11, 4, 11, which after rearranging it in numerical order would be: 3, 4, 5, 5, 7, 9, 9, 10, 11, 11, 11. The measures of central tendency for this data set take on the following values: the mean — 7.73, the median — 9, the mode — 11.

Measures of variability

While measures of central tendency show the most typical score for a set of data, measures of *variability* (or *dispersion*) provide information on the degree to which individual scores vary from the central point in the distribution. The most commonly used measures of variability are the range and the standard deviation.

The *range* is defined as the distance between the highest and the lowest score. It is a useful, first measure of variability. However, the range is largely dependent on extreme scores, which makes it an unstable measure of variability.

The *standard deviation*, formally defined as the square root of *variance* (sum of the squared deviations about the mean divided by N-1), is best understood as similar to, though not equivalent to, the average of the deviations of the scores (ignoring the negative signs), a deviation being the distance of a score from the mean. The standard deviation, therefore, is a statistic that shows how much the scores are spread around the given mean. The larger the standard deviation, the more spread out are the scores (cf. Fitz-Gibbon and Morris 1987).

In the above example of data set, the measures of variability take on the following values: SD — 3.04, Range — 8.

2.5.2. Inferential statistics

Inferential statistics, as has been noted above, are used to help the researcher to make generalizations from the sample to the population under study. The role of this type of statistics, therefore, is to justify inferential claims. To put it simply, using inferential statistics (or *significance testing*, as it is sometimes referred to) allows us to rule out, more or less, the anxiety that a relationship in our sample is just a chance pattern which might not have been there had we happened to look at other samples (Fitz-Gibbon and Morris 1987). The choice of the statistical procedure to establish the amount of confidence you can have in your findings will depend on the number of independent and dependent variables, the number of levels in each variable, type of comparisons to be made in the study, type of data collected, meeting

the assumptions of a particular test, etc. All these procedures, however, are based on the same underlying logic of hypothesis testing. The result that you get is the estimated *probability* that the claims you want to make on the basis of your data are actually wrong. Thus, in statistics, the level of significance is expressed in terms of a decimal fraction where $p<.05$ means that there is less than five per cent probability that your results arose by chance, $p<.01$ means that there is less than one per cent probability that your results arose by chance, etc. (cf. Norton 2009).

In order to estimate the statistical significance of the relationship found in the sample data, the researcher needs to formulate two mutually exclusive hypotheses, which are the statements about the possible outcomes of the study: the *null hypothesis* (H_0), which usually takes the form of the statement of no difference or no relationship, and the *alternative hypothesis* (H_1), which will be adopted if the null hypothesis is rejected, usually it is the same as the research hypothesis. If, for example, a statistical test is used to compare the mean scores obtained in the experimental and control groups, the hypotheses might be formulated as follows:

H_0: There is no difference between the population means of the experimental and control groups ($\mu_E = \mu_C$).
H_1: There is a difference between the population means of the experimental and control groups ($\mu_E \neq \mu_C$).

The alternative hypothesis stated in this way is called a *two-tailed* (or *nondirectional*) hypothesis, as no claims are made as to the direction of the difference. The alternative hypothesis could also be formulated as a *one-tailed* (or *directional*) hypothesis, e.g.:

H_1: The experimental population mean is higher than the control population mean ($\mu_E > \mu_C$),
or
H_2: The experimental population mean is lower than the control population mean ($\mu_E < \mu_C$).

As the next step in hypothesis testing, you need to state the probability level (α-*level*), which is the level of significance at which you will feel confident in rejecting the null hypothesis. In applied linguistics research the α-level is usually set at .05 where there are 5 chances in 100 of being wrong and 95 chances in 100 of being right in rejecting the null hypothesis (Hatch and Lazaraton 1991: 232).

Having set up the probability level, you need to choose an appropriate statistical test. For example, testing the above-listed hypotheses involves

the comparison of means of two independent samples. The appropriate statistical test to apply for this purpose would be the t-test for independent groups or, if the assumptions for the t-test are not met, the Mann-Whitney U test. Irrespective of the procedure, however, after the computer has done its job, you will end up with the *p-value*, which, as has been noted before, is the estimate of the probability of being wrong in rejecting the null hypothesis. Now you need to make a decision if you can safely reject the null hypothesis. If the obtained p-value is lower than the α-level you have set for the study, you can reject the null hypothesis (and accept the alternative hypothesis). If, however, the obtained p-value is higher than the α-level, the null hypothesis cannot be rejected.

While establishing statistical significance of your claims is an important part of quantitative data analysis, your findings still need to be interpreted. As Brown notes, 'the notion of significance does not necessarily imply meaningfulness' (1988: 122). The fact that an observed phenomenon is most probably true in the population and not just in the sample does not necessarily mean that it is important (Dörnyei 2007). Moreover, when drawing conclusions, you should always keep in mind that the numbers you get are just imperfect operationalizations of the constructs under investigation, so they should be approached with some caution. As Norton puts it:

> If you decide to take an experimental approach in your pedagogical action research, the benefits are that you will have a research study where the evidence will be quantitative, and statistical analysis will allow you to interpret the statistical significance of your findings. You will still need to be very careful though in not overgeneralising from your findings, as no matter which basic experimental design you choose, you are not working in a laboratory with inert substances but in the field where educational research with human participants is never straightforward and rarely produces clear-cut findings which cannot be challenged (2009: 106).

3. Overview of sample studies

As noted by Nunan (1992: 91), experimental studies are comparatively rare in classroom research where the data are collected in genuine classrooms, that is, from the intact groups of learners. Since they are time-consuming, labour intensive, and require some expertise in research methodology and quantitative data analysis, teacher-researchers rarely choose this research method. Even if they do, the results of their work seldom get published. More reports of such studies come from professional researchers working in collaboration with practicing teachers. These studies, however, often go beyond what may traditionally be classified as action research. Below I present one experimental research study that, although inspired by theoretical considerations, produced results meaningful to the practice of foreign language teaching and learning. The first study is discussed in considerable detail, the second one is summarized in Table 1.

Study 1

The study was conducted at the University of Granada, Spain, by Diane Naughton (Naughton 2006). The study focused on investigating the effect of a cooperative strategy training program on the patterns of oral interaction with the aim of enhancing small group communication in the classroom. The rationale for the study was the observation that not all communication among learners in the foreign language classroom leads to language learning gains. As noted by Naughton in the introduction to the study, SLA research suggests that in order to profit from communicative interaction learners need to engage in negotiation for meaning (e.g. Long 1996) where communicative misunderstandings trigger some type of feedback (e.g. Pica 1994) leading to noticing the gap between the learner's interlanguage and the target language model (Schmidt 1990). L2 development is also said to be facilitated by learners' engaging in negotiation of form, for instance, when they interact with each other in order to produce collaborative output (e.g. Swain and Lapkin 1998). The role of collaborative effort in language development is also stressed when language learning is viewed from a sociocultural perspective (Vygotsky 1981) where 'linguistic development is seen to emerge through the social mediation of the group's activity' (Naughton 2006: 170). According to Naughton, however, due to certain limitations, the monolingual FL classroom environment does not necessarily encourage interaction patterns

that are conducive to L2 development. Nevertheless, learners can be taught to engage in communicative tasks in ways that enhance language learning. The quasi-experimental study reported on in the article was set up in order to evaluate the effectiveness of the collaborative strategy training program that had been designed to teach the learners how to successfully use the following four strategies: using follow-up questions, requesting and giving clarification, self- or other-repair (recasts), and requesting and giving help. To this end, the following research questions were posed (Naughton 2006: 172):

1. Will the overall interaction patterns that emerge during small group discussion be affected by cooperative strategy training?
2. To what extent do students interact strategically before training and can strategic interaction be further encouraged through the training program?
3. With what frequency do students use the individual strategies before and after exposure to the program?

The study sample was composed of forty-five adult Spanish EFL students coming from five intact classes that were randomly designated as two control groups (N=21) and three experimental groups (N=24). Within each class, the students were subdivided into groups of three. The data on interaction patterns were gathered in two videotaping sessions. The triads were videotaped in a small conference room by their classroom teacher at the beginning (pretest) and the end (posttest) of the eight-week experimental period. All the triads took part in the same pretest and posttest unstructured discussion task (they were given a card with the main topic and possible subtopics to discuss). The overall participation was measured by the number of turns taken. A turn was defined, after Chaudron, as 'any speaker's sequence of utterances bounded by another speaker's speech' (1988: 45). Use of interaction strategies was defined by the number of times the students engaged in asking follow-up questions, requesting and giving clarification, self- and other-repair, and requesting and giving help. The data were gathered with the use of an observation-tally form filled in by the researcher and two independent raters. In the course of the study, the experimental groups received 8 hours of strategy training as a part of their regular 40-hour EFL course. In the control groups, the same time was devoted to unstructured small group discussion work, in which the students were given some topics to discuss.

The data were analysed with the use of descriptive and inferential statistics. As regards the first research question, some evidence was found in support of the hypothesis that cooperative strategy training had an effect on the overall participation in small group communication. Descriptive statistics showed that the mean number of turns taken in the experimental

groups increased between pretest and posttest measures and the mean number of turns taken in the control groups decreased. This finding was confirmed by some, though not all, inferential statistics computed in the study. Naughton's conclusion was that the strategy training program had some effect on the interaction patterns of the experimental group students, though the results should be viewed with some caution. As concerns the second research question, the difference between pretest and posttest scores on total strategic participation was not statistically significant in the control groups ($p>.05$), whereas a pronouced increase between the pretest and posttest was found in the experimental groups ($p<.05$). In Naughton's opinion, it seems that the strategy training led to an increase in overall strategy use, whereas working in unstructured small groups did not. In order to address the third research question, the effect of strategy training on strategy use was analysed separately for all four interaction strategies. The use of follow-up questions did not change in the control groups ($p>.05$), but it significantly rose in the experimental groups ($p<.05$). The same pattern was found for requesting and giving help. As regards requesting and giving clarification, a decrease was observed in the control groups ($p<.05$), and an increase in the experimental groups ($p<.05$). The use of self- and other-repair increased in both the control and experimental groups ($p<.05$). Naughton offers a detailed discussion and interpretation of these results together with their pedagogical implications, which cannot be presented here due to space limitations. Her final conclusions are as follows:

> Cooperative interaction can be aided through the teaching of certain strategies [...] which foster certain types of behaviour and cognitive engagement, as well as metacognitive reflection among students. The teacher should be responsible for modeling strategic interaction and for providing support to the students so that they can progress towards the autonomous use of such strategies. In order to achieve this goal, the learner not only needs to be able to identify learning opportunities, but must also be able and willing to seize them in collaboration with his or her peers, and appropriate elements of socially constructed dialogue for individual cognitive development. In this sense, oral interaction in the monolingual FL classroom may well be a dynamic and powerful source of L2 development (Naughton 2006: 180).

Study 2

The study was conducted in a junior high school in Osaka, Japan, by Tomoko Tode (2007). The design and results of the study are summarized below.

An example of a study based on the quasi-experimental design:

Article:
Tode T., 2007: "Durability problems with explicit instruction in an EFL context: the learning of the English copula *be* before and after the introduction of the auxiliary *be*". *Language Teaching Research* 11, 1, 11—30.
Purpose of the study:
Investigating long-term effects of explicit and implicit grammar instruction.
Main research questions:
1. Do explicit instruction and implicit instruction have positive effects on the learning of the English copula *be* by Japanese junior high school learners?
2. Do the effects of explicit and implicit instruction in (1) hold after the auxiliary *be* is introduced?

Subjects:
89 Japanese junior high school students.
Type of design:
Quasi-experimental (three intact groups: two experimental and one control).
Dependent variable:
Command of English copula *be*;
Operational definition: suppliance of the copula *be* in obligatory contexts on a written discrete-point translation test;
Test: Pretest, two posttests following instruction on the copula *be*, three posttests following the introduction of the auxiliary *be*.
Major independent variable:
Type of instruction: explicit, implicit, control.
Type of analysis:
Descriptive statistics, Inferential statistics (factorial ANOVA).
Results:
Explicit instruction had positive effects in the short term, while implicit instruction did not. The effects of explicit instruction did not last, especially after presentation of the progressive.
Conclusions:
Instruction after explicit instruction must include some adjustments to retain its effect.

4. Questions and tasks

1. What is an experiment? Is the experimental method useful in research on foreign language learning? What are the differences between the experimental studies in traditional SLA research and the experiments conducted within the action research framework? What are the advantages and disadvantages of the experimental method?
2. In Diane Naughton's (2006) study on collaborative strategy training reported in Section 3 above, the subjects were informed that they were being filmed so that the teacher could analyse their interaction patterns; they were not informed, however, of their experimental or control condition. What effect, in your opinion, this information could have had on the outcomes of the study? If the students had been informed of their experimental or control status, how could that have affected the results? In the same study, the observation forms were filled in by the researcher and two independent raters. What was the reason for implementing this procedure?
3. In order to test the hypothesis that attending a language course with a strategy training component (STC) brings better results in language learning than attending the regular language course (RC), three independent experimental studies were conducted using the classic control group pretest-posttest design. In all the studies, the data collected supported the research hypothesis. The results of the three studies were as follows:

Study 1: The group whose language course was preceded by ten hours of training in the use of cognitive, metacognitive and socioaffective strategies obtained significantly higher scores on the First Certificate in English (FCE) exam taken immediately after completing the course.

Study 2: The group whose language course was complemented with five-minute memory strategy training sessions at the beginning of each class got significantly higher grades based on the arithmetic average of the scores on five vocabulary tests taken throughout the semester.

Study 3: The group whose language course included five video presentations illustrating the use of communication strategies in oral interaction had, on average, a lower number of communication breakdowns at the end-of-the-year oral interview.

Were all the researchers equally justified in concluding that incorporating strategy instruction in the language course facilitates language learning? Why? Why not?

4. Read the following abstract (adapted from Abraham 2001) and answer the questions below.

> The present study examined the use of multimedia software for enhancing vocabulary learning and reading comprehension of one hundred and two students enrolled in intermediate-level Spanish classes. Specifically, the research investigated the effects of annotations (glosses) in the form of video, photographs, Spanish definitions, and English (L1) definitions on learning new words and understanding an authentic story in Spanish. A control group did not have access to annotations while reading the story. The choice-lookup group freely looked up annotations for 85 words and the forced-lookup group was required to consult all annotation types available for the 85 annotated words. Students completed a pretest and posttest of 20 Spanish words annotated in the story for which they provided an English translation. In addition, all participants wrote a summary of the story in English and completed a test of their verbal ability in English. Participants in the choice-lookup and forced-lookup groups performed significantly better on the measures of vocabulary learning and reading comprehension than the control group. No statistically significant differences in performance were found between the choice-lookup and forced-lookup groups on the vocabulary posttest and summary. Verbal ability in English did not influence performance on the vocabulary posttest or summary for the three groups.

a) What research questions were posed by the researcher?
b) What research hypotheses were investigated in the study?
c) What variables were included in the design of the study and how were they classified (dependent, independent, moderator, control)? How were they operationally defined?
d) What were the results of the study? Were all the research hypotheses confirmed?
e) What conclusions can be drawn from the study? Can any teaching implications be formulated?

5. In an attempt to test the hypothesis that introducing new vocabulary with the use of multisensory techniques leads to better long-term retention than more conventional techniques of vocabulary presentation, an experimental study has been set up in which two groups of EFL learners subjected to two different types of instruction were asked to write three vocabulary tests: a pretest preceding the treatment, a posttest immediately following the treatment, and a delayed posttest

administered two months after the treatment. Hypothetical results of the study are presented in Figure 4. Interpret these results and answer the following questions:

a) Which of the graphs show(s) the results that confirm the hypothesis posed in the study?
b) Which of the graphs show(s) the results that contradict the hypothesis?
c) Which of the graphs in your opinion shows the most probable results? Why?

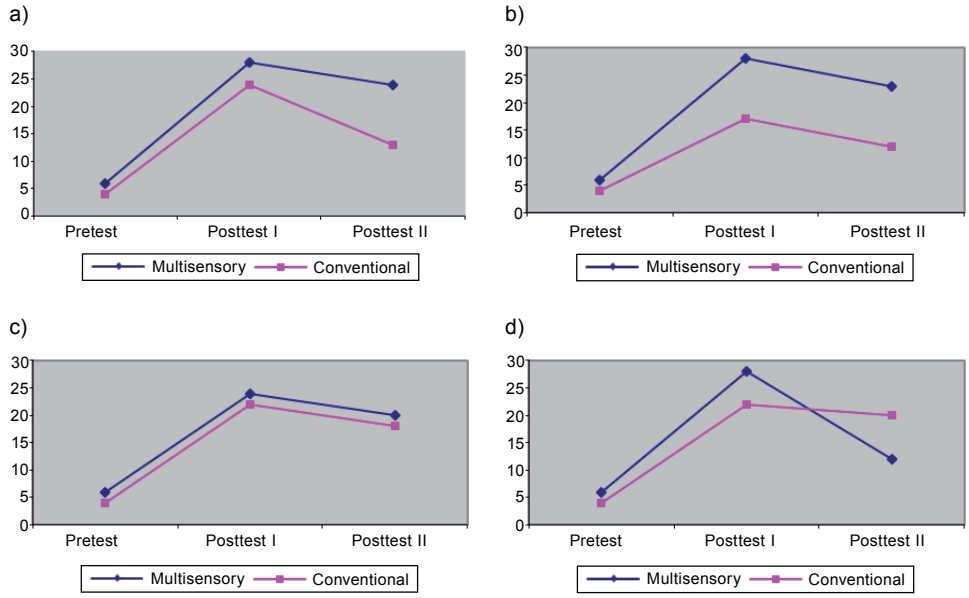


Fig. 4. Hypothetical results of the study of the effectiveness of multisensory techniques in vocabulary presentation

6. The following table contains the hypothetical scores on the listening component of the tests administered in the study presented in Figure 2 (see Section 2.4.3). Study the table, complete the graph presenting the results and answer the questions below. How would you interpret the results?

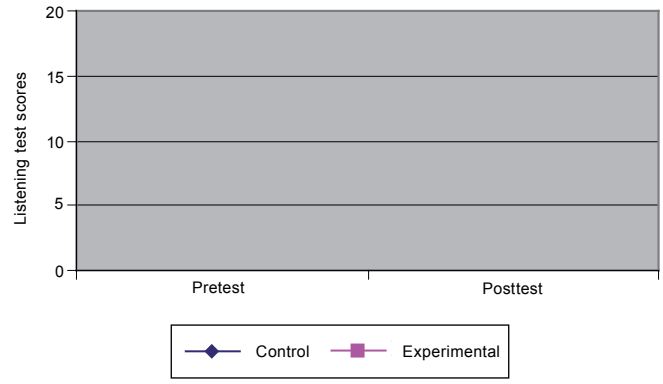
a) Which group, on average, scored higher on the pretest? Was the difference between the pretest means statistically significant?
b) Which group, on average, scored higher on the posttest? Was the difference between the posttest means statistically significant?
c) Did the listening skills of the subjects in the control group improve over the eight weeks of the study? What was the average growth in scores? Was it statistically significant?

d) Did the listening skills of the subjects in the experimental group improve over the eight weeks of the study? What was the average growth in scores? Was it statistically significant?
e) What was the difference in the average gain scores of the control and experimental groups? Was the difference statistically significant?
f) Which group was more heterogeneous with respect to listening test scores? How do you know that?
g) What conclusions can be drawn from the study? Are the findings meaningful?

Table 1. Hypothetical results of Ms Bird's study (see Figure 2, Section 2.4.3) — descriptive statistics

Listening comprehension	Pretest		Posttest		t-test results (dependent samples)	Gain scores	
	mean	SD	mean	SD		mean	SD
Control Group (N=20)	8.25	2.99	13.05	3.24	t (19) = 7.82 p<.0001	4.80	2.74
Experimental Group (N=20)	8.05	3.17	16.85	3.60	t (19) = 9.95 p<.0001	8.80	3.96
t-test results (independent samples)	t (38) = 0.21 p>.05		t (38) = 3.51 p<.01		—	t (38) = 3.72 p<.001	

Fig. 5. Hypothetical results of Ms Bird's study (Complete the graph using data from Table 1)



Additional reading

Brown J.D., 1988: *Understanding Research in Second Language Learning.* Cambridge, UK: Cambridge University Press. A popular introduction to reading statistical research, contains a discussion of essential elements of research design together with the description of the basic statistical tests used in research in second language learning.

Dörnyei Z., 2007: *Research Methods in Applied Linguistics.* Oxford, UK: Oxford University Press. Written in a reader-friendly, accessible style, Dörnyei's book provides a good and thorough overview of different research methods. Chapters on quantitative data collection and analysis will be of particular interest to researchers planning experimental studies. Moreover, his discussion of mixed methods research illustrates how experimental studies can be combined with other methods to generate more meaningful findings.

Francuz P. and Mackiewicz R., 2007: *Liczby nie wiedzą, skąd pochodzą. Przewodnik po metodologii i statystyce nie tylko dla psychologów.* Lublin: Wydawnictwo KUL. A book any student and beginner researcher must read in order to believe that quantitative data analysis can actually be fun. Full of entertaining examples, this book can get anyone to understand statistics. Written by psychologists, it contains a thorough discussion of the experimental design.

Hatch E. and Lazaraton A., 1991: *The Research Manual. Design and Statistics for Applied Linguistics.* Boston: Heinle and Heinle Publishers. Now considered a classic, this manual is indispensable in reading and designing quantitative research studies. Rather conservative in the approach to using statistics, it contains a thorough coverage of parametric and nonparametric statistical procedures useful in quantitative data analysis.

Howell D.C., 1999: *Fundamental Statistics for the Behavioural Sciences.* Pacific Grove, CA: Duxbury Press. A thorough introduction to the use of statistics in the behavioural sciences, full of interesting examples and exercises, written in accessible, at times humorous, language. Focusing on the importance of context and interpreting results, Howell's book shows that there is more to statistical analysis than just applying a few equations.

Norton L.S., 2009: *Action Research in Teaching and Learning. A Practical Guide to Conducting Pedagogical Action Research in Universities.* London and New York: Routledge. A very good and simple introduction to pedagogical action research written for students working on their diploma papers and theses. The chapter on quantitative data analysis contains an overview

of basic statistical terms and guides the students through the process of choosing appropriate statistical procedures.

References

Abraham L.B., 2001: "The effects of multimedia on second language vocabulary learning and reading comprehension". A doctoral dissertation. The University of New Mexico.

Allwright D., 1993: "Integrating 'research' and 'pedagogy': Appropriate criteria and practical possibilities". In: J. Edge and K. Richards (eds) *Teachers Develop Teachers' Research*. London: Heinemann.

Allwright D. and Bailey K., 1991: *Focus on the Language Classroom: An Introduction to Language Classroom Research for Language Teachers.* New York: Cambridge University Press.

Breen M.P., 1985: "The social context for language learning — a neglected situation?". *Studies in Second Language Acquisition* 7, 135—158.

Brown J.D., 1988: *Understanding Research in Second Language Learning.* Cambridge, UK: Cambridge University Press.

Brzeziński J., 2008: *Badania eksperymentalne w psychologii i pedagogice.* Warszawa: Wydawnictwo Naukowe Scholar.

Burns A., 1999: *Collaborative Action Research for English Language Teachers.* Cambridge, UK: Cambridge University Press.

Burns A., 2005: "Action Research". In: E. Hinkel (ed.) *Handbook of Research in Second Language Teaching and Learning.* Mahwah, NJ: Lawrence Erlbaum Associates.

Campbell D.T. and Stanley J.C., 1963: "Experimental and quasi-experimental designs for research on teaching". In: N. Gage (ed.) *Handbook of Research on Teaching.* Chicago: Rand McNally.

Chaudron C., 1988: *Second Language Classrooms: Research on Teaching and Learning.* New York: Cambridge University Press.

Chaudron C., 2005: "Data Collection in SLA Research". In: C.J. Doughty and M. Long (eds) *The Handbook of Second Language Acquisition.* Oxford, UK: Blackwell Publishing.

Crookes G., 1993: "Action research for second language teachers: Going beyond teacher research". *Applied Linguistics* 14, 2, 130—144.

Dörnyei Z., 2007: *Research Methods in Applied Linguistics.* Oxford, UK: Oxford University Press.

Doughty C.J. and Varela E., 1998: "Communicative focus on form". In: C.J. Doughty and J. Williams (eds) *Focus on Form in Classroom Second Language Acquisition*. Cambridge, UK: Cambridge University Press.

Fitz-Gibbon C.T. and Morris L.L., 1987: *How to Analyze Data*. Newbury Park, CA: Sage Publications.

Francuz P. and Mackiewicz R., 2007: *Liczby nie wiedzą, skąd pochodzą. Przewodnik po metodologii i statystyce nie tylko dla psychologów.* Lublin: Wydawnictwo KUL.

Greene J.C. and Caracelli V.J., 2003: "Making paradigmatic sense of mixed methods practice". In: A. Tashakkori and C. Teddlie (eds) *Handbook of Mixed Methods in Social and Behavioural Research.* Thousand Oaks, CA: Sage.

Grotjahn R., 1987: "On the methodological basis of introspective methods". In: C. Faerch and G. Kasper (eds) *Introspection in Second Language Research.* Clevedon: Multilingual Matters.

Harley B., 1989: "Functional grammar in french immersion: a classroom experiment". *Applied Linguistics* 10, 3, 331-359.

Hatch E. and Lazaraton A., 1991: *The Research Manual. Design and Statistics for Applied Linguistics.* Boston: Heinle and Heinle Publishers.

Howell D.C., 1999: *Fundamental Statistics for the Behavioural Sciences.* Pacific Grove, CA: Duxbury Press.

Johnson R.B. and Onwuegbuzie A.J., 2004: "Mixed methods research: a research paradigm whose time has come". *Educational Researcher* 33, 7, 14—26.

Johnson R.B. and Christensen L., 2004: *Educational Research: Quantitative, Qualitative, and Mixed Approaches.* Boston: Allyn and Bacon.

Komorowska H., 1982: *Metody badań empirycznych w glottodydaktyce.* Warszawa: Państwowe Wydawnictwo Naukowe.

Larsen-Freeman D. and Long M., 1991: *An Introduction to Second Language Research.* Harlow: Longman.

Long M., 1996: "The role of the linguistic environment in second language acquisition". In: W.C. Ritchie and T.K. Bhatia (eds) *Handbook of Language Acquisition: Vol. 2 Second Language Acquisition.* New York: Academic Press, 413—468

Long M., 2007: *Problems in SLA.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mackey A., 1999: "Input, interaction and second language development". *Studies in Second Language Acquisition* 21, 4, 557—587.

Mertens D.M., 2005: *Research and Evaluation in Education and Psychology: Integrating Diversity with Quantitative, Qualitative, and Mixed Methods.* 2nd Edition. Thousand Oaks, CA: Sage.

Naughton D., 2006: "Cooperative strategy training and oral interaction: enhancing small group communication in the language classroom". *The Modern Language Journal* 90, 2, 169—184.

Norton L.S., 2009: *Action Research in Teaching and Learning. A Practical Guide to Conducting Pedagogical Action Research in Universities.* London and New York: Routledge.

Nunan D., 1992: *Research Methods in Language Learning.* Cambridge, UK: Cambridge University Press.

Nunan D., 2005: "Classroom research". In: E. Hinkel (ed.) *Handbook of Research in Second Language Teaching and Learning.* Mahwah, NJ Lawrence Erlbaum Associates.

Ortega L. and Long M., 1997: "The effects of models and recasts on the acquisition of object topicalization and adverb placement in L2 Spanish". *Spanish Applied Linguistics* 1, 1, 65—86.

Pica T., 1994: "Questions from the language classroom: research perspectives". *TESOL Quarterly* 28: 49—79.

Roberts J., 1998: *Language Teacher Education.* London: Edward Arnold.

Schmidt R., 1990: "The role of consciousness in second language learning". *Applied Linguistics* 11, 129—158.

Stevens S.S., 1951: "Mathmatics, measurement, and psychophysics". In: S.S. Stevens (ed.) *Handbook of Experimental Psychology.* New York: John Wiley.

Swain M.K. and Lapkin S., 1998: "Interaction and second language learning: two adolescent French immersion students working together". *Modern Language Journal* 82, 320—338.

Tode T., 2007: "Durability problems with explicit instruction in an EFL context: the learning of the English copula *be* before and after the introduction of the auxiliary *be*". *Language Teaching Research* 11, 1, 11—30.

Van Lier L., 1988: *The Classroom and the Language Learner.* London: Longman.
Van Lier L., 1996: *Interaction in the Language Curriculum.* London: Longman.
Vygotsky L.S., 1981: "The genesis of higher mental functions". In: J.V. Wertsch (ed.) *The Concept of Activity in Soviet Psychology.* Armonk, NY: M.E. Sharpe.
Wallace M., 1998: *Action Research for Language Teachers.* Cambridge: Cambridge University Press.

Joanna Bielska

Zastosowanie eksperymentu w badaniach w działaniu

Streszczenie

Rozdział podejmuje temat wykorzystania metody eksperymentalnej w badaniach nad uczeniem się i nauczaniem języka obcego, skupiając się na przydatności tej metody dla prowadzenia tzw. badań w działaniu (ang. *action research*).

Metoda eksperymentalna przedstawiona została tu w kontekście debaty pomiędzy jakościowym a ilościowym paradygmatem badawczym. Nacisk kładziony jest również na różnice pomiędzy zastosowaniem metody eksperymentalnej w badaniach nad przyswajaniem języka obcego, w których służy ona do weryfikacji hipotez w celu rozwijania teorii, a jej wykorzystaniem w badaniach nad optymalizacją procesu nauczania języka obcego, których wyniki mają przede wszystkim zastosowanie praktyczne.

W rozdziale omówiono podstawowe cele i założenia metody eksperymentalnej a także wprowadzono terminologię niezbędną do poprawnego stosowania tej metody. Przedstawione zostały najczęściej stosowane rodzaje eksperymentów, zasady ich przygotowywania oraz opisano podstawowe metody statystycznej analizy danych. Wykorzystanie metody eksperymentalnej zostało następnie zobrazowane kilkoma przykładach zaczerpniętych z literatury przedmiotu. Umieszczone na końcu rozdziału pytania i zadania przygotowują czytelnika do krytycznego odbioru tekstów zawierających raporty z badań eksperymentalnych, mogą one być również przydatne przy samodzielnym planowaniu eksperymentów.

Joanna Bielska

Die Anwendung eines Experimentes in Aktionsuntersuchungen

Zusammenfassung

Zum Gegenstand des Kapitels wird die Ausnutzung der experimentellen Methode in den Forschungen über die Fremdspracheerlernung und Fremdsprachenunterricht und besonders deren Anwendung bei sog. Aktionsuntersuchungen (eng. *action research*).

Die experimentelle Methode wird hier im Zusammenhang mit der Diskussion über das qualitative und quantitative Forschungsparadigma dargestellt. Betont werden auch Unterschiede zwischen der Anwendung des Experimentes in den Forschungen über die

Fremdspracheerlernung, wo sie zur Verifikation von Hypothesen zur Theorieentwicklung dient, und deren Ausnutzung in den Forschungen über die Optimierung des Fremdsprachenunterrichts, deren Ergebnisse hauptsächlich praktische Anwendung haben.

In dem Kapitel werden Hauptzwecke und Hauptthesen der experimentellen Methode und die deren richtigen Anwendung dienende Terminologie besprochen. Es ist hier die Rede von den am häufigsten angewandten Experimenten, deren Vorbereitung und von den wichtigsten Methoden der statistischen Datenanalyse. Die Anwendung der experimentellen Methode ist an Hand einiger der Fachliteratur entnommener Beispiele veranschaulicht. Die sich am Ende befindenden Fragen und Aufgaben sollen den Leser auf kritische Rezeption der Forschungsberichte vorbereiten und beim selbständigen Planen des Experimentes helfen.