CHAPTER SIX

Experimental Research

Susan Gass

Experimental research is a way of determining the effect of something on something else. In other words, a researcher begins with an idea of why something happens and manipulates at least one variable, controls others, to determine the effect on some other variable. To take a simple example from everyday life, let's suppose that an individual has developed a rash, but the source of the rash is not known. There are a number of possibilities: (1) something eaten, (2) a medication taken, (3) touching an object that causes an allergic reaction. We might eliminate one factor at a time, let's say, a particular medication taken to see if the rash disappears. If it does not, then one eliminates a medication taken as the source of the problem. This continues until the culprit is discovered. There is, of course, a fallacy in this reasoning because the rash could disappear on its own, given the many unknown workings of the human body. Nonetheless, experimental research, in this case, is a way of determining, to the extent possible, what the source of the rash is.

We turn to an example from language. Let's assume that we want to know whether focusing a learner's attention on some aspect of language increases that individual's learning of that aspect of language. One way to do this is to find two groups of learners which are matched on their pre-experiment knowledge of this aspect of language. There is then a treatment session in which the experimental group receives focused attention on the particular part of language under investigation, while the control group receives exposure to the same part of language, but their attention is not intentionally directed. A post-test measures improvement from the pre-test. In sum, experimental research involves the manipulation of at least one

variable, known as an independent variable, while keeping other relevant variables constant, and observing the effect of the manipulation on some other variable (known as a dependent variable), for example, a test score.

Underlying assumptions of the methodology

Even though a discussion of the philosophy of science is beyond the scope of this chapter, it is important to at least acknowledge some of the philosophical underpinnings when considering what experimental research is and why some individuals are 'married' to this approach and others do not see its value. In the history of second language acquisition (SLA) research, there are those who tend to be what might be called rationalists in their approach to research and others who are considered relativists in their approach (see Jordan 2005 for a disentanglement of these and other related terms). In the former approach, one finds a strong role for experimental quantitative research. The latter group sees the value of multiple realities and more frequently uses qualitative data; in other words, there is no objective reality – only our perceptions of reality.[1] These divisions in recent years are becoming blurred with many studies using both quantitative and qualitative data. These are known as mixed methods studies (see Hashemi & Babahii 2013 for a review; see also the *Journal of Mixed Methods Research*).

Larsen-Freeman and Long (1991) discuss two approaches to research that are differentiated by whether they might be considered inductive or deductive. Conducting research and then coming up with a theory (research-then-theory) is inductive, whereas a theory-then-research approach is deductive. Experimental research with its rationalist underpinnings falls into this latter category.

Specifics of experimental research

With experimental research, there are a number of components that researchers include. First, there is a specific and precise research question. This research question follows from some theoretical vantage point and is generally based on something that remains unanswered from previous research. Second is a set of explicitly stated variables, that is, what is being varied and what is being measured. These must reflect the construct under investigation. Third is a randomly selected group of participants who are randomly assigned to various treatment conditions and/or control groups. Additionally, in reporting research results, the treatment must be described explicitly and must relate in a direct way to the research question. Interpretation of the results, that is, of the effects of the treatment is generally done through statistical means. Finally, to ensure that the interpretation of

the treatment effects is accurate, there must be controls and there must be appropriate counterbalancing. In the following section, we elaborate on each of these parts.

Research questions

Research questions must be stated explicitly and must have some basis in previous literature. For example, consider the following two examples:

1 Does focused attention on noun–adjective agreement in Italian promote learning to a greater extent than focused attention on *wh*-movement in Italian for beginning learners of Italian?

Prior research (Gass, Svetics & Lemelin 2003) argued that focused attention promoted learning in some parts of the grammar and not in others and differentially affected learners at different proficiency levels.

2 Should language classes be introduced early in a school's curriculum?

The first question is explicitly stated, although as we will see below, there are variables that are in need of further elaboration and greater explicitness. The second question, while interesting and one which school districts are constantly debating, is not researchable given the vagueness as well as the word *should* which implies some sort of right or wrong and which, as a result, cannot be empirically evaluated.

In sum, in experimental research, there need to be answerable questions. There are a number of characteristics that one can think of when determining whether a question is answerable or researchable. Given the limited resources (time and money) of most researchers, questions must be feasible in relation to the time and budget of the problem. This often entails scaling back on the level of complexity of the research question. As is implied from the discussion above, the question needs to be significant in relation to current research in the field of inquiry. And, of course, ethics must play a role in the questions that are asked. This topic is dealt with in greater detail below. Question 2 above fails in that it would probably not be feasible given the never-ending number of variables involved, even though it is significant in that it is an important question for all school districts.

Research hypothesis

In experimental research, there are not only research questions, but also hypotheses. Hypotheses are predictions based on the research question. For example, in research question 1 above, the following hypothesis might obtain.

1 Focused attention on noun–adjective agreement in Italian will promote learning to a greater extent than focused attention on *wh*-movement in Italian for beginning learners of Italian.

Alternatively, if there is no reason to expect anything other than a difference, the hypothesis might be phrased as follows:

2 Focused attention on noun–adjective agreement in Italian will promote learning to a different degree than focused attention on *wh*-questions in Italian for beginning learners of Italian.

Often, these are expressed as *null* hypotheses which state that there are no differences between/among groups. The research goal is to reject the null hypothesis.

Variables

Variables are characteristics of a class of objects that vary, as in the variable of eye colour or height or weight for humans. In experimental research, variables need to be made explicit. This is actually one of the most difficult parts of a research project. To take research question 1 above, there are a few variables about which decisions need to be made to render those variables sufficiently explicit, so that a researcher can conduct his/her research. For example, how will focused attention be operationalized (by colouring instances of noun–adjective agreement in a written text? By providing an explicit grammatical description? By frequency – that is by introducing numerous instances of noun–adjective agreement in a passage)? When conducting and reporting experimental research, researchers must be clear on how they are defining their terms.

A second variable in the research question under consideration is the notion of learning. In the history of SLA, different definitions have been used and here too numerous decisions have to be made. Is learning measured by a paper/pencil (or computer) test or by spontaneous use? If by a paper/pencil or computer test, what kind of test? If the latter, how does one elicit spontaneous use? And, what does it mean if a form is not used? Does it mean that the learner does not know it or does it mean that the learner does not think she or he needs to use it? If the latter, then learning may have taken place as a result of the treatment, but there may be extraneous reasons for it not being used. Further, if spontaneous use is the criterion for learning, how many instances of the structure/sound/lexical item need to occur? Mackey (1999), for example, required 'the presence of at least two examples of structures in *two* different posttests' (p. 567). Is learning being operationalized only through a test given immediately following the treatment, or will the test be given one week later, or a month later? Other

variables relate specifically to the question being considered. For example, investigating noun–adjective agreement in Italian requires further decisions. Are only feminine nouns to be included, only masculine nouns or both? Because nouns that end in *-a* or in *-o* almost invariably indicate what the gender of the noun is, will nouns that do not obviously indicate gender (e.g. *ponte*, m. bridge) be included? Similarly, for questions (*wh*-movement), one will have to determine which questions to include; does one include questions that include prepositions (e.g. *for whom*, *with what*)?

In sum, variables are characteristics that vary. In experimental research, there are essentially two primary variables of concern: independent variables and dependent variables. Independent variables are the object of investigation. They are those variables that the researcher is investigating in order to determine their effect on something else. In the above example, focused attention on the two grammatical structures is the independent variable. Dependent variables are those variables that the independent variable is having an effect on. Thus, learning is the dependent variable in the above example. In addition to these variables are extraneous variables; these variables are independent variables that the researcher has not controlled for. In essence, they can seriously interfere with the results of a study. In the above example of focused attention, let's assume that a researcher has operationalized focused attention by colouring red all instances of noun–adjective agreement in a reading passage. Let's also assume that the researcher did not test for possible colour-blindness of the participants. Then, colour-blindness is an uncontrolled variable that could have interfered with the interpretation of the results.

Random assignment

A third hallmark of experimental research is the random assignment of participants to one group or another. Random assignment of individuals means that each individual has an equal chance of being assigned to any of the conditions of the study (experimental or control). That is, the process of assignment is random. Randomization is intended to eliminate the possibility that extraneous variables will creep into the research design. To take the above example, had there been randomization, there would have been an equal chance of having colour-blind individuals in both the adjective–noun group and the *wh*-question group.

In general, in educational settings, we cannot always have random assignment of individuals and we are more often dependent on the contexts that already exist (e.g. intact classes) for our research. This is known as quasi-experimental research (as opposed to true experimental research) because not all variables can be completely controlled; in particular, we are dependent on assignment of participants based on class placement rather

than on random assignment. But, in such instances, there should be random assignment of the group/class to one condition or another.

Statistical analysis

Interpretation of results is done as a first step through statistical analyses. A discussion of different kinds of statistics and designs is beyond the scope of this chapter, but there are numerous books that deal with these topics (see Dörnyei 2007; Mackey & Gass 2005, 2012, 2015). In general, for most research in the field of SLA or applied linguistics more generally, the significance level (α level) is generally set at 0.05. This means that when we look at statistical results, there is a 95 per cent chance that our results are due to the experimental treatment and only 5 per cent chance that the results are due to chance alone. In many studies, researchers will talk about the results *approaching significance* when the statistical analysis is slightly about 0.05 (e.g. 0.06 – 0.09). In disciplines where the consequences of a chance finding are greater (e.g. life-altering medical treatment), different α levels are required. In other words, the α level is a generally accepted guide that is used by researchers in a discipline. When we are accepting or rejecting hypotheses, we want to avoid any errors in interpretation. There are two noteworthy error types, known as Type I (also known as an α error) and Type II (also known as a β error). The former refers to the rejection of a (null) hypothesis when it should not be rejected, and the latter refers to the acceptance of a (null) hypothesis when it should be rejected. Both of these errors are minimized through rigorous and appropriate use of statistics.

Statistical significance gives us a 'yes/no' indication of the significance of one's results. In most cases, a dichotomous decision is insufficient; one wants to know instead how strong findings are. Effect sizes, because they are not dependent on sample size, are frequently used (many journals require an effect size as part of reporting). Probability levels are highly dependent on sample size, whereas effect sizes are not. Plonsky and Oswald (2014) provide a full discussion of effect sizes and their interpretation.

As part of the interpretation process, one has to be able to eliminate with a reasonable degree of confidence that other factors did not enter into the picture. There are ways to minimize this possibility. One, we have already discussed and that is randomization. Another was hinted at with the colour-blindness example and that is by testing for that variable and eliminating those participants who have that characteristic. One could also include that variable into the design by including it as a variable and then testing for its influence. And, a variation of the latter is to match participants, so that one has a particular characteristic and another does not. This latter is perhaps better understood if one took the variable of gender. Let's assume that in the focused-attention experiment we have

reason to believe that males might behave differently than females, we then would want to balance males and females in all groups.

Validity and reliability

With all research, we need to be able to be confident in our results; our results need to be trustworthy and valid. Validity encompasses many of the concepts already discussed in this chapter. In the following section, we refer to a concept that is often discussed together with validity and that is reliability. *Validity* refers to the correctness and appropriateness of the interpretations that a researcher makes of his/her study. Reliability refers to score consistency across administrations of one's instrument. For both concepts, accurate and appropriate instruments are at the core. Thus, if our hypothesis involves learning, we need to have an accurate and appropriate instrument to measure learning. For example, if we are looking at knowledge representation, we need to have a measure that appropriately reflects that and not something that just measures an ability to use the language. In many instances (see discussions in Gass & Mackey 2007; Gass with Behney & Plonsky 2013), part of the theoretical discussions in the literature involve how best to represent particular constructs.

Often variables cannot be measured directly. In these instances, we come up with a working definition that allows us to identify the variable in question with something that is understandable and measurable. This is known as an operationalization. Thus, in our example presented earlier of focused attention, we cannot directly measure this construct, but we can come up with a reasonable surrogate (e.g. colouring or highlighting in some way). Once we have operationalized a variable, we can more easily work with it.

The last thing that a researcher wants is to spend time, effort and money on a project and then realize that the study itself did not reflect what we had thought it would and might apply only to the population of the study and not to the broader community at large. Validity comes in many different colours; in this section, we discuss the most common types of validity: content, face, construct, criterion-related and predictive validity. Following that brief introduction, we turn to a discussion of internal and external validity.

Content validity

Content validity refers to the representativeness of our measurement regarding the phenomena that we want information about. If, for example, we want information about noun–adjective agreement in Italian, we cannot

generalize these findings to say that we have fully investigated all types of agreement (article–noun, singular–plural, regular–irregular, nouns that end in *–a-o* and those that are not morphologically marked as masculine/feminine). In other words, if we want to claim that we have investigated agreement, we need to ensure that our instruments include a representative range of what constitutes agreement.

Face validity

Face validity is closely related to the notion of content validity and takes us into the realm of the consumers of research. Is our instrument readily recognizable as measuring what we claim it measures? For example, the construct of intelligence can be measured in various ways, but there are certain instruments that are well-accepted as measuring this construct, even if it is a somewhat elusive construct. Thus, face validity refers to the familiarity of our instrument, and how easy it is to convince others that there is content validity to it. If a school district wants to measure intelligence with a newly developed instrument, there may be a perception by the community (e.g. parents) that this instrument is not valid (unless of course their child receives a high score!). If the participants do not perceive a connection between the research activities and other educational or second language activities, they may be less likely to take the experiment seriously.

Construct validity

Construct validity refers to the extent to which the research adequately captures the concept in question. In second language and applied linguistic research, construct validity is of great concern because a great deal of what we investigate is not easily quantifiable and not directly measurable. Some variables, such as height, weight and shoe size, are easily measureable and there is little controversy over what they reflect. Thus, a weight of fifty-two kilos is clear to everyone who uses this scale, but in second language research, we are dealing with such constructs as proficiency. What precisely does this mean? How can we measure it, so that we can compare individuals on a common scale? Because these constructs are not directly measurable, their validity can be called into question. One way to enhance construct validity is to have multiple measures. Thus, if we were to measure proficiency, we might have measures that reflect oral use, written use, extent of vocabulary knowledge and so forth. If we were to use these measures as an aggregate, we could have greater confidence in our ability to differentiate individuals along a scale of proficiency.

Criterion-related validity

Criterion-related validity refers to the relationship that a given measure has with some other well-established measure. For example, if a researcher develops an overall measure of language proficiency, it will have criterion-related validity if it measures language students in much the same way as another well-established test. To be more specific, if we are doing a study using first-, second- and third-year English learners of Spanish and develop a test that measures oral proficiency, criterion-related validity would be increased if we could show that on our test, third-year students did better than second-year students who did better than first-year students. Our test, then, would correspond to some other reasonably accepted measure, that of class placement.

Predictive validity deals with how well the measure we are using predicts performance on some other measure. In other words, if we have a test that measures working memory capacity, it has predictive validity if it predicts performance in class performance in a language class.

Internal and external validity

In addition to these five types of validity, there are two other types of validity that are noteworthy: internal validity and external validity. Each of these is important when conducting experimental research.

To what extent are the results of a study truly reflective of what we believe they reflect? This is known as *internal validity*. In other words, are our dependent and independent variables related in the way we think they are? A researcher must control for (i.e. rule out) all other possible factors that could potentially account for the results. This was discussed in the example above in relation to colour-blindness. In that study, had we not controlled for colour-blindness, we would have been left with the unfortunate conclusion that the study had little internal validity. Before conducting any research study, we need to think carefully through the design to ensure that we eliminate or minimize threats to internal validity (see Mackey & Gass 2005, 2015 for a more thorough treatment of this topic and for ways to minimize threats to internal validity).

External validity refers to the potential generalizability[2] of a study. We can make conclusions about the behaviour of the participants in a study, but this is not particularly interesting unless the results have broader implications and are relevant to a wider range of language learners and language learning contexts. Thus, if we conduct a study with English-speaking learners of Italian studying at University ABC, we are interested not just in those specific learners, but also the extent to which the results are applicable to learners of other languages possibly in different contexts. This

is the case because we are interested in general principles of learning and not just a particular group of learners.

External validity can be increased with appropriate sampling procedures, as mentioned above. In particular, it is important that our sample be selected randomly, which, in essence, means that each member of the population to be studied has an equal and independent chance of being selected. This is the ideal situation, but one which, in reality, is not always practical. Rather, in second language research, nonrandom sampling is frequently used. Researchers often seek volunteers to participate in a study, as is required by most university ethics review boards. Even when intact classes are used, students can opt out of participation according to university ethical requirements. Sufficiently large sample sizes are always a goal as a way to increase the likelihood of true differences between groups (e.g. experimental and control). Small sample sizes leave the researcher with the uncertainty of interpreting the results. Are the differences between groups true or just coincidental? Many statistical tests help researchers avoid drawing unwarranted conclusions.

Because true random sampling is not always likely in second language research, it is important for researchers to fully and accurately describe the population studied as well as provide details about the materials, methods and procedures. In this way, a particular study can be replicated by others which, in a way, broadens the population base of the original study (see Polio & Gass 1997; Porte 2012 for a fuller description and discussion of replication and reporting). Mackey and Gass (2005, 2015) provide additional discussion on issues of external validity and outline ways of minimizing threats to external validity.

Reliability

Reliability refers to consistency and is a way of ensuring that our constructs are being measured appropriately. In applied linguistics research, it is frequently used when raters are making judgements about data. This is referred to as interrater reliability (when more than one rater is involved) and intrarater reliability (when only one researcher's evaluations are used). In the former instance (e.g. judging oral speech samples on a scale of 1–10), consistency across raters indicates that raters are measuring the same construct in the same way. In the latter case, one might rate the same speech sample at two different points in time to ensure consistency.

Both validity and reliability are ways of ensuring quality in research. As noted, experimental research is a way of finding answers to questions in a disciplined way. These results may have far-reaching impact (including decisions relating to educational practices), and it is incumbent on the

research community to ensure that research (experimental and other) is carried out in as careful a way as possible, ensuring quality at each step of the way.

Techniques and instruments

There are as many techniques and instruments as there are research projects. A description of even a few would go beyond the scope and appropriate length of this chapter. Two recent books (Ellis & Barkhuizen 2005; Gass & Mackey 2007) deal with data elicitation methods and methods of analysis respectively and can serve as useful sources of information.

Ethical considerations

As with all research, ethical considerations abound. The most obvious concerns the protection of human subjects and will not be dealt with here (see Mackey & Gass 2005, 2015, Chapter 2; see also De Costa, in press, this volume) as they are elaborated on by each institution's ethical review board. In brief, in most educational settings, one must obtain permission from a human research committee before conducting any research or before recruiting volunteers for a research project (see https://www.citiprogram.org/citidocuments/forms/Responsible%20Conduct%20of%20Research%20(RCR)%20Catalog.pdf for information on responsible conduct of research, viewed 26 January 2015). The overriding concern is that no harm comes to participants with the ideal being that there be benefits.

As researchers design studies, there is often a control group against which to measure the effects of a particular treatment. But, here too, there is an ethical question. If we have reason to believe that our treatment is beneficial, then we have recruited volunteers who will not receive the treatment. One possibility is to provide the treatment to the control group after all data have been collected. Polio and Gass (2007) designed a study in just this way, although even in this study, there were limitations to equal treatment. The main research question was: Can a brief intervention study in which pre-service teachers are instructed on how to interact with learners promote learning in a subsequent interaction? The design involved an experimental and a control group. The experimental group received a 15–20 minute session with a researcher who went over ways to increase student output (e.g. asking open-ended questions rather than yes/no questions). This was followed by an interactive session with an ESL learner in which they were asked to put into practice what they had learned. So as not to disadvantage

the control group and maintain the integrity of the study, the training session with the control group was conducted following the interactive session. Thus, all pre-service teachers had the benefit of a training session which had been hypothesized to be beneficial in the promotion of learning. Yet, the control group's training was conducted after the experiment so as not to influence the research results. This, of course, does not take into account the benefits that the ESL students in the experimental group had over those in the control group. Unfortunately, the complete balancing of benefits would have required the control group of ESL students to return for a second round, and this was not logistically possible. In general, it is important to strictly follow the guidelines established by one's institution regarding all aspects of a study, including modes of recruiting, actual treatment and assessment details and the reporting of information in such a way as to respect privacy and anonymity issues. In addition to local review boards, the American Psychological Association (2010) has important guidelines for many aspects of the research process.

A final point to be made which has ethical ramifications has to do with honesty in reporting, and in particular with the elimination of participants. When elimination takes place, it has to be done judiciously and with justification. For example, in studies which measure reaction time, it may be the case that a participant is not focused on the task. This may be determined by inordinately fast reaction times which make it clear that she or he is just pushing a response button without processing the required material. Often, in such studies, a cut-off point will be determined, such as two standard deviations[3] above or below the mean, with individuals falling on the outside limits of this cut-off point being eliminated from analysis. Whatever criteria are used, it is important that a detailed and principled justification be provided to avoid any question of impropriety.

A sample study

The study that I have chosen to highlight is a study by Alison Mackey, Susan Gass and Kim McDonough, 'How do learners perceive interactional feedback' published in 2000 in *Studies in Second Language Acquisition*. This study was selected because it is an empirical quantitative study, bolstered with qualitative data. The study illustrates the problem of a possible uncontrolled variable cropping up in the study which in this case led to a further study. Thus, it illustrates the cycle of conducting a study, analysing the results (including a post-hoc analysis – an analysis that did not result from the research questions that guided the study) and postulating an uncontrolled variable that led to a further quantitative study (Gass & Lewis 2007).

Description and research questions

As stated in Mackey et al. (2000, p. 477)

> The focus of the current study is an exploration of the claim that, through negotiated interaction, learners' attention may be directed toward particular aspects of language. In order to explore whether interactional feedback and the allocation of focal attention to feedback play a role in the development of L2 knowledge, it is important to first investigate the extent to which that feedback is in fact perceived as such by learners and whether their perceptions about the target of the feedback are correct.

The specific research question was: how do learners perceive the feedback they receive in the course of interaction?

Participants

There were seventeen ESL learners and eleven Italian as a foreign language (IFL) learners in this study.

Task

Each learner carried out an interactive task with a speaker of the language they were learning (English or Italian). When there were errors, the researcher provided corrective feedback to the learner. Following the interaction (which had been videotaped), the video was replayed to the learner who, using a Stimulated Recall procedure (see Gass & Mackey 2000), was asked to comment on what they were thinking during the moments of feedback. These tapes were coded by two raters with the rating being based on the feedback given and the perception of that feedback. Interrater reliability was calculated and reported. There were four (excluding those that were unclassifiable or had no classifiable content) categories of feedback: (1) phonological, (2) morphosyntactic, (3) lexical and (4) semantic. The responses were then paired according to whether the intent of the feedback corresponded with the perception of the feedback. In other words, was phonological feedback perceived as phonological feedback?

The study did not have a control group–experimental group design; rather, it dealt with static perception, that is, perception of an event. Results were presented in terms of percentages (no statistics in the main study) of feedback of phonology, of morphosyntax and of lexis (semantics had too few responses to comment on).

There are times when the data from a study reveal possibilities of interpretation that had not been planned for from the outset. In this study, there were two post-hoc analyses that were conducted as well as a suggestion that led to a further study.

The post-hoc analyses first analysed the type of feedback (recast, negotiation or a combination) in relation to the error type. This analysis was necessary (but not planned) because the interactions had not been scripted and the feedback was naturally occurring. The second post-hoc analysis related learners' perceptions about feedback and their immediate uptake.

Results

The results of this study showed that learners perceive different error types differentially, that there was a different distribution of feedback type depending on error type and that uptake was different depending on error type. Interestingly, there were differences between the ESL group and the IFL group. It was speculated that the difference might relate to the heritage population[4] that had been included in the IFL group.

Further research

Much research leaves as much unanswered as answered. This research was no exception and, in fact, spawned a different study (Gass & Lewis 2007) which dealt with the variable that 'snuck' into the study (also known as an intervening variable), and which may have influenced the results, that of heritage versus non-heritage learners.

Notes

1 Other terms that are used are constructvist/interpretivists who primarily use qualitative research methods and positivists or empiricists who rely primarily on quantitative methods. In reality, research today crosses both camps and it is not uncommon – yours truly included – to find researchers using both quantitative (experimental) and qualitative methods in their quest to answer research questions.

2 There are instances where generalizability may not be a goal of a study. This might be the case within a particular instructional context in which curricular changes are being debated. In such instances, there is a need to determine the extent to which a proposed curricular change (e.g. inserting a technological component into the curriculum) results in increased learning, although in the case of a technological component, the goal might be 'no decrease in learning' because the purpose behind a technological component in a curriculum might be financial.

3 A standard deviation is a measure of dispersion. It is a numerical value that indicates how the scores in a sample are spread out in relation to the mean and therefore indicates the homogeneity or lack thereof of a sample.
4 Heritage learners are those who come to the learning situation with some degree of exposure to the language being learned through their family background. There is a wide range of learner profiles within this category, ranging from the target language being spoken exclusively as a home language to only infrequent use of that language, as for example, visiting relatives abroad during summer visits.

Resources for further reading

Brown, JD & Rodgers, T 2002, *Doing Second Language Research*, Oxford University Press, Oxford.

This book covers a range of research topics including qualitative (case studies, introspective data, classroom data) and quantitative research. In the latter category, there is a discussion of descriptive statistics, correlational research and quasi-experimental research. Within these categories is information about compiling and analysing data as well as guidelines for designing and interpreting data.

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

This text covers a range of research types (quantitative, qualitative, longitudinal). Included are discussions of theoretical and philosophical underpinnings of research types. Basic issues such as data collection and analysis are discussed, as are guidelines for reporting research.

Gass, S & Mackey, A 2007, *Data Elicitation for Second and Foreign Language Research*, Lawrence Erlbaum, Mahwah, NJ.

This book is an extension of Mackey and Gass' (2005) book on research methods (see below). The book focuses extensively on ways of collecting second language data. Chapters, organized around research approaches (e.g. psycholinguistics, formal approaches, interaction), include a discussion of research questions and historical underpinnings of some of the techniques. Each research approach is exemplified with data elicitation techniques including naturalistic language, prompted linguistic production and non-linguistic experimental responses.

Mackey, A & Gass, S 2015, *Second Language Research: Methodology and Design*, 2nd edn, Routledge, New York, NY.

This book, targeted towards students of SLA, addresses basic issues related to research design, providing step-by-step instructions for how to carry out studies. Topics include identifying research problems and questions; selecting elicitation measures; dealing with ethical issues related to data gathering; validity and reliability in research; research in classroom contexts; qualitative research, mixed methods research, data description and coding; and data analysis. Also included is a chapter on writing research reports with suggestions about preparing research results for publication.

Mackey, A & Gass, S (eds), 2012, *Research Methods in Second Language Acquisition: A Practical Guide*, Wiley-Blackwell, Malden, MA.

This edited collection consists of fifteen chapters (including an introductory chapter by the editors), each of which deals with the *how* of *a* particular area of second language research (e.g. reading, writing) or of particular types of data (e.g. corpus, classroom, survey, case study, qualitative, psycholinguistic, formal theory-based). In addition, the second part of the book provides practical information about how to code quantitative and qualitative data, how to conduct appropriate statistical analyses and how to conduct meta-analyses. Finally, there is a paper that focuses on the *why*, *when* and *how of replication studies*.

Porte, G 2002, *Appraising Research in Second Language Learning: A Practical Approach to Critical Analysis of Quantitative Research*, John Benjamins, Amsterdam.

Porte's book focuses on understanding and interpreting research reports. The goal of the book is to produce critical readers of research. The book is organized around research reports, namely, the abstract, introduction, review of literature, participants, materials, procedures, results, discussion and conclusion. Questions are peppered throughout each of the sections leading the reader to a critical understanding of appropriate content for a sound research article.

Porte, G (ed.), 2012, *Replication Research in Applied Linguistics*, Cambridge University Press, Cambridge.

Porte's edited volume deals specifically with replication research. It consists of nine chapters written by scholars in the field and introductory and concluding chapters by Porte. The book focuses on definitions of replication, arguments for the importance of replication, practical considerations when conducting replication studies as part of graduate programs and in practice. Porte's forward-looking final chapter focuses on the numerous challenges involved in conducting and publishing replication studies, not the least of which is the role of journals and other publication venues in the process.

References

American Psychological Association 2010, *Publication Manual of the American Psychological Association*, 6th edn, American Psychological Association, Washington, D.C.

De Costa, P (ed.), in press, *Ethics in Applied Linguistics Research: Language Researcher Narratives*, Routledge, New York, NY.

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

Ellis, R & Barkhuizen, G 2005, *Analyzing Learner Language*, Oxford University Press, Oxford.

Gass, S with Behney, J & Plonsky, L 2013, *Second Language Acquisition: An Introductory Course*, 4th edn, Routledge, New York, NY.

Gass, S & Lewis, K 2007, 'Perceptions of interactional feedback: Differences between heritage language learners and non-heritage language learners', in A Mackey (ed.), *Conversational Interaction in Second Language Acquisition: A Series of Empirical Studies*, Oxford University Press, Oxford, pp. 173–196.
Gass, S & Mackey, A 2000, *Stimulated Recall Methodology in Second Language Research*, Lawrence Erlbaum Associates, Mahwah, NJ.
——— 2007, *Data Elicitation for Second and Foreign Language Research*, Lawrence Erlbaum, Mahwah, NJ.
Gass, S, Svetics, I & Lemelin, S 2003, 'Differential effects of attention', *Language Learning*, vol. 53, no. 3, pp. 497–545.
Hashemi, M & Babaii, E (eds), 2013, 'Mixed methods research: Toward new research designs in applied linguistics', *The Modern Language Journal*, vol. 97, no. 4, pp. 828–852.
Jordan, G 2005, *Theory Construction in Second Language Acquisition*, John Benjamins, Amsterdam.
Larsen-Freeman, D & Long, M 1991, *An Introduction to Second Language Acquisition Research*, Longman, London.
Mackey, A 1999, 'Input, interaction and second language development', *Studies in Second Language Acquisition*, vol. 21, no. 4, pp. 557–587.
Mackey, A & Gass, S 2005, *Second Language Research: Methodology and Design*, Lawrence Erlbaum, Mahwah, NJ.
——— 2012, *Research Methods in Second Language Acquisition: A Practical Guide*, Wiley-Blackwell, Malden, MA.
——— 2015, *Second Language Research: Methodology and Design*, 2nd edn, Routledge, New York, NY.
Mackey, A, Gass, S & McDonough, K 2000, 'How do learners perceive interactional feedback?', *Studies in Second Language Acquisition*, vol. 22, no. 4, pp. 471–497.
Plonsky, L & Oswald, F 2014, 'How big is "big"? Interpreting effect sizes in L2 research', *Language Learning*, vol. 64, no. 4, pp. 878–912.
Polio, C & Gass, S 1997, 'Replication and reporting', *Studies in Second Language Acquisition*, vol. 19, no. 4, pp. 499–508.
——— 2007, 'Getting students to talk: preservice teacher intervention and learner output', Paper presented at the Fifth International Conference on Language Teacher Education, University of Minnesota.
Porte, G (ed.), 2012, *Replication Research in Applied Linguistics*, Cambridge University Press, Cambridge.