CHAPTER TWO

Quantitative Research and Analysis

Aek Phakiti

A good deal of research in applied linguistics and language acquisition research involves quantifying aspects of language learning and use and factors that are part of or influence language learning and/or use (e.g. knowledge about language, first language, motivation, self-regulation and anxiety (see Lightbown & Spada 2013; Macaro 2010; Ortega 2009 for topics of language learning research). Lazaraton (2005), Loewen and Gass 2009, Plonsky and Gass (2011) and Plonsky (2013, 2014) present and discuss issues and applications of quantitative methods and designs in applied linguistics and language learning research. This chapter introduces quantitative research methods in language learning and use and discusses issues involved in quantitative data analysis. It also presents two types of statistical analyses (descriptive and inferential statistics) that are used in quantitative research. The chapter concludes with the presentation of a sample study, paying particular attention to quantitative data analysis.

Underlying assumptions and methodology

Quantitative research usually involves numbers, quantification and statistics to address a research problem or objective, and it typically requires a large sample size. There are two kinds of statistics which determine the nature of quantitative research: descriptive statistics and inferential statistics (discussed below). Quantitative research that aims to investigate the characteristics of a population (e.g. Census survey, a

national literacy test) or opinions, perceptions and attitudes of learners (e.g. perceived effectiveness of a language program, learning anxiety and language needs) may simply aim to report an *average score*, a *percentage* or a *rank* of something for all participants. This kind of quantitative research uses *descriptive statistics*.

Another type of quantitative research aims to do more than just reporting an average score, ranked score or a percentage. For example, researchers may aim to examine a causal-like or linear relationship between two or more variables, such as aspects of language learning (e.g. the age, gender, language proficiency and/or aptitude of students and the teaching methods employed to teach them). This kind of quantitative research uses *inferential statistics*.

The origin of quantitative research was *positivism* – a philosophical perspective that strives to understand universal principles or rules that govern human behaviours. A positivist takes a *realist stance* and believes that reality is essentially independent of who is examining it and when and how it is being examined. Today's quantitative researchers generally take a *postpositivistic* position, which is a modified version of positivism. It distinguishes ideology from reality. According to proponents of this position, an object of inquiry cannot be understood with complete accuracy. Instead, they believe that objectivity is a *regulative* principle which reminds researchers to be aware of potential influences of their personal bias or values as well as those of others involved in a research setting on research outcomes. See Guba and Lincoln (2005) and Phakiti and Paltridge (this volume) for a discussion of research paradigms.

Types of quantitative research

Examples of research where a quantitative approach is usually taken are experiments, surveys, correlational studies, individual differences research and language testing research. It should be noted that although today's research may be qualitative or quantitative or mixed, in this chapter each of these designs is discussed in the situation in which researchers employ a quantitative method.

Experimental research

Experimental researchers consider two types of variables (i.e. aspects or characteristics of something that can take different values or scores): independent and dependent variables. An *independent variable* is a variable that exists freely and is hypothesized to have an effect on other variables that are described as *dependent variables*. For example, if an instruction is an independent variable, students' learning performance

in the classroom may be considered a dependent variable. Experimental research typically examines causal-like relationships between independent variables and dependent variables of interest. Instructed second language acquisition research can test a theory using an experimental research design, which should be considered and planned in its entirety prior to its implementation. Researchers strictly control the conditions under which an independent variable of interest (e.g. types of feedback or instructions) is to be tested for its effects on language learning behaviours or outcomes (e.g. better writing performance and increased positive attitude towards learning). *Random assignment* is used to allocate participants into experimental and control groups. Random assignment is a technique in which researchers assign participants into two or more groups using a by-chance technique (e.g. a coin-toss or randomized computer program). Through this technique, each participant will have an equal chance of being placed in any one group (see Gass this volume). Inferential statistics, such as *t*-tests and analysis of variance (ANOVA), are used to answer research questions (see e.g. Gass this volume; Phakiti 2014; Read this volume; Vandergrift this volume). In recent years, experimental research has begun to utilize a mixed methods approach.

Survey research

Survey research often focuses on a snapshot of a particular topic of interest (i.e. it is cross-sectional) with a large sample size. It can be longitudinal when researchers aim to collect the same data from the same participants at different points of time. Researchers typically adopt *Likert-scale questionnaires* which ask respondents to choose a scale ranging, for example, from 1 to 5, or *structured interview*s which allow all respondents to answer the same questions. It is important to note that survey research is not always purely quantitative, because researchers can gather qualitative data (e.g. through open-ended questions or survey interviews) and analyse them qualitatively. Survey research can, therefore, be quantitative, qualitative or mixed, depending on a researcher (see Wagner this volume). When survey researchers analyse quantitative questionnaire data, they typically use statistics to examine the reliability of the questionnaire, correlations between pairs of variables of interest and to make comparisons among different groups of participants.

Correlational research

Correlational research aims to explore whether a relationship between two variables exists, and if so, to what extent they are related. It is conducted without manipulating the research setting. Survey research can adopt

correlational analysis, but correlational research is broader than survey research in that the data are not always gathered from survey instruments. Data may be from an existing database (e.g. students' grades, test scores and hours of study) or participants may be asked to complete test tasks so that their test performance data can be used for correlational analysis. Research that focuses on a particular language skill (e.g. reading, listening, speaking and writing) also uses correlational analysis (see several chapters in Part 2 of this volume).

Individual differences research

Individual differences research (also known as *ex post facto* research) examines whether a variable (e.g. motivation, aptitude, anxiety, language learning strategies, reading ability and attitude towards something) differs among different groups of learners (e.g. defined by gender, age, years of language exposure, language background and native versus non-native speakers). Data may be collected using similar methods to those used in correlational analysis or using survey research instruments. As in experimental research, groups of learners are treated as independent variables for the purpose of statistical comparison. However, unlike in experimental research, researchers do not manipulate, modify or control the research setting. They explore differences among learners as they exist, as in survey and correlational research.

Language testing and assessment research

Language testing is a key area of research in applied linguistics. It aims to measure language learners' proficiency or language skills. The number of language testing studies is quantitative because it examines test reliability, scoring methods and test validity (see below). Test validation research that aims to verify and improve the quality of test questions or tasks before its actual use in testing language learners is predominant in this type of quantitative research. Several specific statistical procedures, such as Item Response Theory, Generalizability Theory and Differential Item Functioning analysis are used (see Purpura 2011; Read this volume for further discussion of this).

Validity and reliability

There are two interconnected levels of validity in quantitative research: the validity of a research instrument and the validity of a research result. The *validity of a research instrument* (e.g. a test, a questionnaire and an

observation scheme) is related to how accurately the instrument yields information about the aspect under investigation. An evaluation of the validity of a research instrument includes, for example, considering the theory underlying the behaviours to be measured, the characteristics of the instruments (e.g. types and number of questions or items), the manner in which it is used to collect the data and its reliability (i.e. consistency of the scoring, rating or reporting).

Reliability is related to the concept of *consistency*. To understand the concept of reliability, let us take a clock as an example. A clock with a good battery is likely to be consistent, thereby reliable. Suppose one clock is five minutes faster and the other is ten minutes slower than the actual time. Both clocks are consistent in their timing. In this example, the clocks are reliable, but they do not tell us the correct (let us say 'valid') time. The concept of reliability is closely related to validity in the sense that a quantitative study cannot be valid if it uses unreliable data to analyse and answer its research questions. It is important to note that something that is valid in one context may not be valid in another. Take a watch as another example. Three students have just arrived in Sydney (Australia) from Shanghai (China), Barcelona (Spain) and Tokyo (Japan). They have not adjusted their watches yet. Their watches' times are not valid to show Sydney's time but are valid in their home countries.

In quantitative research, the reliability of the data is often determined by reliability coefficients or measures of internal consistency. In language test data, researchers usually report KR-20 (Kuder–Richardson 20) as a reliability estimate of objective tests (Brown 1996). In rating-scale and Likert-scale questionnaires, researchers report a correlation coefficient or Cronbach Alpha coefficient (Dörnyei 2007). Generally speaking, a reliability estimate ranges from 0 (0 per cent reliable) to 1 (100 per cent reliable). A reliability coefficient of 0.90 upwards is desirable, although a coefficient of 1 is rare. In general, a reliability coefficient of 0.70 for a questionnaire is taken to be acceptable (Dörnyei 2007).

The *validity of a research finding* is concerned with the *accuracy* of the finding. This validity depends on how valid and reliable research instruments in a particular study are as well as on the statistical validity of the data analysis. *Statistical validity* is related to the *probability value* that researchers set to reject a *null hypothesis* (a null hypothesis states that no relationship or difference among groups exists). Statistical validity also depends on whether researchers use an appropriate statistical test to investigate a research question and whether the statistical assumptions of a particular statistical test are met or not. The considerations of both the validity of the research instruments being used and the research outcomes reflect the extent to which a quantitative study is *sound* and is likely to yield useful information or knowledge (see Gass this volume, which presents specific aspects of quantitative validity including construct, content and face validity).

Techniques and instruments

There is a wide range of quantitative data types in applied linguistics research. Quantitative data can derive from measurement instruments that quantify variables and factors, such as the results of language tests or responses on Likert-scale questionnaires. They can also derive from qualitative data collection techniques such as think-aloud protocols, retrospective interviews, diaries and written, visual, audio or spoken texts (see Gass & Mackey 2007; Mackey & Gass 2005). Several chapters in this volume present and discuss quantitative research instruments and techniques in detail (e.g. Hyland, Stevenson, Vandergrift, Wagner).

Key stages in quantitative data analysis

There are a number of stages researchers typically move through in the preparation of quantitative data for analysis.

Checking and organizing data

In the checking and organizing data stage, the data need to be checked to see whether each study participant has fully completed all sections or items in the data collection procedure. Interview data, for example, need to be transcribed and organized prior to quantification processes (such as tally and frequency counts). Once all data (quantitative or qualitative) have been double-checked, participant identification (ID) codes need to be assigned, so that the data can be easily referred to when checking the accuracy of the data entry.

Coding data

The aim of data coding is to classify or group the data, thereby making it easier to analyse. Coding allows the researcher to quantify and represent the area under investigation using numbers. To achieve validity in the data coding, the data needs to be coded in a systematic and principled way.

How quantitative data is coded depends on the nature of the scales used to measure the variables. There are three basic types of quantitative data, *nominal data*, *ordinal data* and *interval data* (Dörnyei 2007). The distinctions among these are important when selecting appropriate statistical techniques to use to analyse the data. *Nominal data* (also referred to as categorical data) are used for classification and group comparison purposes.

We may, for example, ask participants to report whether they are native or non-native speakers of English. For statistical analysis, we can code this data by assigning 1 to native speakers of English and 2 to non-native speakers of English. Nominal data do not have mathematical properties. For example, we cannot say that people assigned 2 have higher scores than people assigned 1.

Ordinal data (meaning 'ordering' the data) are known as *rank-ordered* data. Examples of this include academic grades, Likert-scale data, rank of class achievement and ranking on the basis of personality or psychological traits. For example, academic grades can be coded as follows: A (coded 5), B (coded 4), C (coded 3), D (coded 2) and F (coded 1). Although ordering data in this way allows us to express differences in individuals' characteristics, it does not allow us to express the degree of these differences. For example, in an agreement scale ranging from 1 (strongly disagree) to 5 (strongly agree), we can see that one is greater or lesser than another in terms of level of agreement, but we cannot say that the distance between the scores of 4 (agree) and 5 (strongly agree) is the same as the distance between the scores of 1 (strongly disagree) and 2 (disagree), for example. Whether ordinal data can be seen as interval data (discussed next) has been an ongoing debate among statisticians and quantitative researchers.

Interval data are data measured on an interval scale, which is a scale of measurement in which the distance between any two adjacent units of measurement is the same. Examples of interval data include language test scores, the number of years of study and age. Interval data are suitable for inferential statistics (discussed below) because they have continuous values.

Coding qualitative data, such as think-aloud, written, oral or interview data for quantitative data analysis involves the development of coding systems to quantify the variables of interest. Much of the coding of qualitative data is subjective in the sense that each coder or rater may interpret the meaning of the qualitative data differently. Hence, we need to be mindful of the consistency of coding for each individual coder or rater (intra-coder or -rater reliability) and between two or more coders or raters (inter-coder or -rater reliability; see Mackey & Gass 2005). Data coding training and moderation of coding need to be implemented to achieve consistency. There are some inter-coder reliability estimates that can be used, including: (1) percentage agreement (the ratio of the number of items coded the same to the total number of items coded) and (2) *Cohen's kappa* (a statistic indicating the average rate of agreement for an entire set of data, taking into account both agreement and disagreement between coders). Cohen's kappa is preferred over percentage agreement as it takes into consideration the random chance of coding agreement and the quantity of data coded. A Cohen's kappa coefficient ranges

between 0 and 1. A Cohen's kappa of 0.75 or above indicates a good to very good level of the agreement between two coders (see Altman 1991; Cohen 1960; Fleiss 1981).

Entering data into a computer program

Once the data have been coded and numerical values have been assigned, the data can be keyed into a statistical software program such as IBM SPSS (Statistical Package for Social Sciences), which allows us to handle quantitative data from a large sample. Issues relevant to data entry include naming data files, defining variables for data recording and entering data into a designated file. The data entry process needs to be managed to reduce the chance of errors (see Dörnyei 2007). During data entry, the issues of missing data and potential outliers (i.e. extreme cases which can distort statistical results) will need to be dealt with.

Screening and cleaning data

The screening and cleaning stage concerns the accuracy of data entry and involves the use of a decision-making process for dealing with missing data and incorrect data entry. For example, SPSS can be used to compute a minimum and maximum value for each variable. If, for example, the maximum value of an item is 5 but SPSS reports it as 55 instead, it means that there is a data entry error for this item. We then need to check where the mistake occurred in the data file and correct it.

Data screening can also be performed through the use of diagrams, such as histograms or charts that can be produced by SPSS. Through a visual inspection of these diagrams, impossible values in the data set can be detected. Data screening allows us to identify and correct data entry problems and remove outliers.

Analysing the reliability of data

Quantitative researchers need to make sure that their data are derived from reliable instruments or measures. Reliability is a necessary yet insufficient condition for validity because a study cannot be valid if its instruments are not reliable. As discussed above, the reliability of an instrument is related to its *consistency* in capturing the target construct of the investigation across points in time and its ability to discriminate among participants who possess different levels of the construct of interest of the research (e.g. writing ability, motivation levels). SPSS can analyse research instrument reliability.

Reducing data

It is often the case that there are numerous variables in the same data file for analysis (e.g. test score variables, strategy use items and motivation items). In this case, we will experience difficulty in managing and analysing the data, so we should seek to reduce the number of variables to a reasonable level. If we do a correlational analysis (discussed in the sample study), there can be hundreds of correlation coefficients to interpret, report and discuss. Hence, we need some theoretical rationale to help us reduce the number of variables for quantitative data analysis to answer our research questions. Many psychological and cognitive constructs, such as language proficiency, metacognition, motivation and anxiety, can be inferred through observation of various behaviours or thoughts. Therefore, we can use multiple items to assess one construct. For example, we may be able to understand metacognitive strategy use by devising questionnaire items measuring planning, monitoring and evaluating strategy use. Within each of these subscales, there are a number of items which can be combined to form a single score (or a *composite*). This is done to ensure that the construct of interest is not under-represented by our measurements. By aggregating items, we gain richer information about the construct under examination.

There are different statistical techniques that can minimize the chance of aggregating irrelevant or problematic items with reliable ones. First, *correlation coefficients* (discussed below) can inform us as to whether the items measure the same construct. Here, we should expect a strong correlation among them. Second, reliability analysis in SPSS can help us decide whether items are suitable for inclusion. In *Cronbach alpha analysis*, SPSS allows us to check whether the reliability of each subscale will increase or decrease if we exclude a particular item in the subscale. This information can help us decide whether we should include a particular item to form a composite. Third, we can employ *exploratory factor analysis* (EFA) to help us reduce the number of items for further inferential statistics.

Performing inferential statistics

Quantitative researchers choose appropriate statistical tests (e.g. parametric or non-parametric, discussed below) to answer the research question. Researchers need to know which statistical test can yield the answer to a particular research question. For example, if they aim to address a linear relationship, a Pearson correlation may be suitable, if the statistical assumptions are met. If they aim to examine group differences, an independent-samples *t*-test, an ANOVA or a multivariate analysis of variance (MANOVA) may be employed. Some of these statistical tests will be introduced below.

Basic statistical concepts

Much of what we do in quantitative data analysis involves *statistics* because we do not always have the information we need from the entire population of interest. If the data are from all members of the population, we deal with *parameters*. However, in most circumstances, such as in a classroom context, we only have access to a sample of the target population. In order to estimate the parameters of the population from the sample, we use inferential statistics. Sometimes in our research, we need only basic statistics such as frequency counts, percentages or means (average scores) to answer our research questions. Here we use *descriptive statistics* (i.e. statistics for describing, summarizing and explaining the distribution of a data set). At other times, we need to go through detailed statistical procedures to answer our research questions because of the complex nature of the issue we are researching and the nature of the research. Here, we work with inferential statistics (i.e. statistics for making inferences about population parameters). We now explore the differences between descriptive and inferential statistics and how they are related in quantitative research.

Descriptive statistics

Descriptive statistics are used to describe individual variables. A mean score is an example of a descriptive statistic. Instead of using raw test data to show how students performed in an exit test, we can calculate the mean score which shows how well the students did in the test as a group. Descriptive statistics can be divided into measures of frequency, central tendency and dispersion.

Measures of frequency are mostly used with nominal data. Frequencies can be represented using a histogram, a sector graph or a bar chart. Frequencies can also be represented in a table with percentages calculated cumulatively. *Measures of central tendency* provide an overall picture of the data. The three common measures of central tendency are the mean, the median and the mode. The *mean* is the most common indicator of central tendency. It can be calculated by summing all the scores of a variable in the data set and dividing the result by the number of scores. For example, the mean of the data set 1, 2, 3, 4, 5 is 3 (i.e. 15 ÷ 5). The *median* is the value that divides the data set exactly into two parts of equal size, the data being arranged in ascending order. In the data set described above, the median is also 3. The median can be more appropriate as a descriptor if the mean is distorted by an outlier. For the data set 6, 7, 8, 9, 11, 76, the mean is 19.5 while the median is 8.5. 76 is an example of an outlier. The *mode* is the value that occurs most frequently in the data set. In the data set 12, 13, 14, 14, 14, 16, 18, the mode is 14.

Measures of dispersion describe the *variability* of the data away from the measure of central tendency. However, measures of central tendency do not necessarily tell the entire story of the data. For example, a mean of 24.66 could be achieved in a *homogenous* data set, such as 24, 24, 24, 25, 25, 26, or by a *heterogeneous* data set, such as 2, 7, 15, 28, 46, 50. The mean also depends on the sample size and hence we cannot always compare the means of two different learner groups if the sample sizes differ greatly. Hence, we may need other measures of dispersion. The *standard deviation* is a statistic that describes the variability of the data. The standard deviation indicates by how much the data varies from the mean. A low standard deviation implies the data points may be clustered around the mean, while a high one shows the data points are more spread out away from the mean.

Inferential statistics

Inferential statistics are used when we try to connect individual variables in terms of their relationships. Inferential statistics help us make inferences about population parameters. To perform inferential statistics, however, there are some other statistical concepts we need to understand, including the normal distribution, probability and parametric and non-parametric tests (see Brown 1991, 1992 for an accessible discussion of these terms).

Normal distribution

Whether a data set has a normal distribution or not is related to the spread of the data points around the mean. A normal distribution has a bell-shaped figure. In a perfect normal distribution, the mean, median and mode have equal values. Variables that determine whether a data set has a normal distribution include sample size and the range of scores. Hatch and Lazaraton (1991) suggest a minimum of 30 participants for quantitative data analysis. However, in general, the more data points used, the greater the stability of the data distribution. It is also important that the data set exhibits a range of scores. This is perhaps why interval data are desirable for statistical analysis. A dichotomous variable that takes on values such as yes (coded 1) or no answer (coded 2) cannot lead to a normal distribution regardless of sample size. However, some sets of ordinal scores which appear continuous due to their large sample sizes may have a normal distribution.

SPSS can produce a histogram that may indicate the presence of a normal distribution. Perfect distribution is scarce in applied linguistics data, so we need some criteria to help us determine if the distribution is considered acceptably normally distributed. There are two statistical measures which SPSS can produce to help us decide this. The first is the *skewness statistic*, which tells us the extent to which a score distribution deviates from perfect symmetry (i.e. mode = median = mean). A negative skewed value suggests

that the distribution is skewed towards the right (i.e. mode > median > mean), while a positive value suggests that the distribution is skewed towards the left (i.e. mode < median < mean). The *kurtosis statistic* is related to the peakedness of a distribution (i.e. whether it is flat or sharp). A kurtosis value of 0 suggests that the data set is normally distributed. A negative kurtosis statistic suggests that the distribution tends to be flat, whereas a positive kurtosis statistic suggests that the distribution is peaked. Conservatively, values of skewness and kurtosis statistics within *plus and minus 1* suggest that the data set is acceptably normally distributed (see Carr 2008 for an accessible discussion of this through the use of Microsoft Excel).

Probability and significance values

Probability is related to the degree to which the statistical finding occurs by chance (i.e. due to random variation). This is related to *statistical validity*, which asks whether or not the statistical finding is true or incidental (i.e. found by chance). The *p*-value (*p* = probability) is the likelihood that we will be wrong in the statistical inferences that we make from the data (i.e. when we reject the null hypothesis). $p < 0.05$ (i.e. there are 5 in 100 chances of being wrong) or $p < 0.01$ (i.e. there are 1 in 100 chances of being wrong) are commonly used or found in applied linguistics research.

To clarify the difference between a probability value and a significance value, we need to remember that the significance value will be *fixed* (e.g. the researcher specifies that it must be less than or equal to 0.05 or 0.01). The *probability value*, on the other hand, is *data-driven* and produced by the test statistics. For example, when we set a *p*-value to be less than 0.01 and when a *p*-value of 0.04 is obtained from the data, this data-driven *p*-value is *not statistically significant* at 0.01 because 0.04 is larger than 0.01. However, if we set a probability value at 0.05, the obtained *p*-value is statistically significant at 0.05 because 0.04 is smaller than 0.05. It is important to note that the word *significant(ce)* is not the same as the word *important(ce)* that we normally use. The word *significant(ce)* in quantitative research is associated to hypothesis testing.

Effect sizes

The statement in the Publication Manual of the American Psychological Association (APA 2010) points out that statistical significance *p-values* are not acceptable indices of effect. In several situations, researchers may find a statistical significance, but the finding yields little meaning, leading to *no theoretical or pedagogical practicality*, thereby not always worthy in a practical sense. An effect size is a *magnitude-of-effect estimate* that is independent of sample size (see Ortega this volume). A magnitude-of-effect estimate highlights the distinction between *statistical* and *practical*

significance (see the Sample Study). Larson-Hall (2010, pp. 118–119) provides a table of effect sizes, their formulas and interpretations).

Parametric or non-parametric tests

There are two types of inferential statistics. These are *parametric* and *non-parametric statistics*. The distinction between the two lies in the different sets of statistical assumptions (i.e. preconditions essential for accurate applications of a statistical test) that must be met before the statistical analysis can be undertaken. These assumptions are *not optional*, and if they are not met, there is a heightened risk of making a false inference/claim (e.g. in rejecting the null hypothesis). The main assumptions for parametric tests are (1) the data is normally distributed, (2) the data is interval or continuous and (3) data scores are independent across measures.

A non-parametric test (i.e. a distribution-free test) is suitable for the analysis of frequency data or data that does not meet the normal distribution assumption. A non-parametric test can analyse discrete variables or ranked-order data. Although parametric tests are preferable in quantitative research, non-parametric tests are important for applied linguistics research because some data are not interval or continuous. Furthermore, in some cases in which the data is categorical or dichotomous (e.g. pass or fail scores), we cannot employ parametric tests because the normal distribution assumption. For example, we may want to see if gender, with male learners (coded 1) and female learners (coded 0), is relevant to passing (coded 1) or failing (coded 0) an English test. With this kind of data, there can be a connection between two different dichotomous variables, and a non-parametric test will be required to detect it.

Statistical tests

Although there are various reliable statistical programs available for use in applied linguistics (e.g. Microsoft Excel, IBM SPSS and Statistical Analysis System (SAS)), it is important to note that researchers need to know and understand the logic behind statistical analysis and the standards for a particular statistical test. There is a wide range of statistical tests that are used in applied linguistics research. Some common statistical tests are introduced here.

Correlations

Correlational analysis is used to examine systematic relationships between two variables. The nature of the data (e.g. continuous or categorical data)

determines the kind of correlational analysis that can be used. Examples of correlational tests include Pearson Product Moment correlations, Spearman Rho correlation, Phi correlations and Point–Biserial correlations (see Phakiti 2014). A correlation test is expressed on a scale from 0 to 1. If two variables are strongly correlated, this means that an increase or decrease in one variable will be accompanied by an increase or decrease in the other variable. If two variables are uncorrelated (i.e. 0), there is no systematic relationship between them. A positive (+) correlation suggests a positive association between two variables (i.e. the two variables are associated and move in the same direction in a systematic way). A negative (–) correlation suggests a negative relationship (i.e. the two variables are associated and move systematically in opposite directions).

Factor analysis

Factor analysis is related to the correlational approach because it is used to determine how the observed variables from questionnaires or tests (responses to items) are linked to a common factor which underlies observed behaviours. In quantitative research, an underlying factor of language learner behaviours (e.g. motivation, strategy use and anxiety) is hypothetically assumed to influence how individuals answer questionnaire items. This hypothetical assumption can be tested empirically through the use of factor analysis. There are two types of factor analysis. The first is known as EFA, which is used to explore the clustering of questionnaire items (i.e. to explore whether the items are relatively homogeneous or highly correlated). EFA is useful when researchers are not certain about a particular construct, especially in a new research area. EFA can help researchers reduce the data and can be used during a pilot study, so that researchers can develop a more rigorous research instrument.

The second type of factor analysis is *confirmatory factor analysis* (CFA). CFA is used when researchers are confident that a particular construct underlies a set of questionnaire or test items. CFA is closely connected to quantitative research that employs a structural equation modelling (SEM) approach, which allows researchers to confirm a theory or hypothesis empirically (see Phakiti 2007; Ockey 2014; Woodrow this volume).

Regression analysis

Regression analysis is also an extension of the correlational analytical method. It is used to examine the prediction of one dependent, continuous-scale variable (e.g. reading comprehension scores) based on values of another one or more independent variables (either categorical or continuous in nature; e.g. genders, age groups, motivation). *Simple regression* uses only

one independent variable, whereas *multiple regression* uses two or more independent variables that are correlated with each other, to predict a dependent variable. In a multiple regression, researchers can identify the independent variable that is the best prediction of a dependent variable. A regression coefficient, which ranges from 0 to 1, tells us the extent to which a dependent variable can be predicted, given a one unit change in an independent variable.

Chi-square test

A *chi-square test* is a non-parametric test that indicates whether a relationship between two categorical variables exists statistically. For example, male and female language learners may differ in their preferred choice of English language learning activity (e.g. reading versus speaking). In a contingency table (often referred to as a *cross-tabulation*), a row represents categories of the gender variable and a column represents categories of the subject variable. This table can be constructed by using frequency counts. A chi-square test can inform researchers whether male students are more likely than their female counterparts to choose a particular English language learning activity.

T-tests

There are two types of *t-test* that are used in applied linguistics: a repeated-measures *t*-test and an independent-samples *t*-test. A *paired-samples t-test* examines whether *two* mean scores from the same group of participants differ significantly. For example, we may want to see whether learners' attitudes to writing feedback have changed after a two-month period. We will have pre- and post-instruction questionnaires investigating their attitudes. An *independent-samples t-test* is used to determine whether the mean scores between two groups of students are significantly different.

Analysis of variance

Analysis of variance (ANOVA) has a similar logic to the *t*-tests mentioned above. A *within-group ANOVA* is similar to a paired-samples *t*-test and a *between-groups ANOVA* is similar to an *independent-samples t-test*. The key difference is that ANOVAs can compare more than two group mean scores or levels of an independent variable. A within-group ANOVA can compare the mean scores among pre-, post- and delay-post tests. A between-groups ANOVA can compare three or more groups of participants (e.g. high-ability, intermediate-ability and low-ability) in terms of their self-

regulation. When more than two means are compared in ANOVAs, a *post-hoc* test will be used to identify exactly which groups significantly differ from each other. That is, the ANOVA can flag a statistical significance, but it does not indicate where the mean difference lies.

Ethical considerations

Different types of quantitative research have different ethical considerations due to their differing research aims and designs. For example, experimental researchers need to carefully consider the potential negative impact of their treatments on participants (see e.g. Gass this volume; De Costa this volume). Generally speaking, quantitative researchers (as with all researchers) need to follow ethical protocols to safeguard their research participants in terms of *confidentiality* and their *right to privacy*. Participants have the right to know what is involved in the research and what they will be doing in the study. They need to voluntarily agree to take part. In survey research, *anonymity* is important because it encourages participants to be truthful in expressing their thoughts and attitudes. Researchers can protect their participants' identities through the use of *pseudonyms* (made-up names) to refer to them and the research site (e.g. school, college, university or company) in the study. Researchers should submit their research ethics application forms to the relevant research ethics review committees in their institution, and it should be approved prior to the data collection.

A sample study

Phakiti (2003) illustrates quantitative data analysis. This study investigated the nature of cognitive and metacognitive strategy use and its relationship to English as a foreign language (EFL) reading test performance. Three hundred and eighty-four Thai university students participated in the study. The study employed an 85-item, multiple-choice reading test to assess the students' English reading achievement on a course in which they were enrolled and a 35-item, 5-point Likert-scale strategy use questionnaire to measure their cognitive and metacognitive strategy use (e.g. 0 = never to 5 = always). The participants first took a 3-hour reading test, which was followed by the completion of the questionnaire. For the purpose of data matching, the questionnaire was put together with the test answer sheet. In the first part of the questionnaire, participants were asked to provide demographic information, including their gender, age and the number of years they had been learning English. In the second part, they were asked to report on the extent of their strategy use during the test completion.

Data preparation

In order to key the test score data and strategy use data into SPSS, each participant was assigned an ID code, so that the data could be checked at a later stage. With regard to the questionnaire data, these were coded according to different types of data as discussed above. For example, for the data relating to gender, males were coded as 1 and females as 2. All this data were then keyed into SPSS. Two people took turns in entering the data to reduce the chance of an error being made. On average, it took about two and a half minutes to enter one set of questionnaire responses into SPSS. The researcher double-checked for accuracy in the data entry at the end by scanning the data file as well as by randomly checking the data against the actual questionnaires.

After the completion of the data entry, descriptive statistics were calculated, including the mean, median, mode, maximum, minimum and skewness and kurtosis statistics of the reading test and questionnaire data. This process allowed the researcher to check whether there were missing data or outliers in the data and to determine whether the data exhibited a normal distribution. Because it was important to see if the data in the reading test and questionnaire were reliable, a reliability analysis was performed. The reliability estimate of the overall test (like KR-20) was 0.88. The Cronbach alpha coefficients (as discussed earlier) for the cognitive strategy and metacognitive strategy variables were 0.75 and 0.85, respectively.

Data analysis and results

In Phakiti's study, three research questions were asked. However, for the purpose of this chapter, only research question 2 will be discussed. This research question was: what is the relationship of cognitive and metacognitive strategies to EFL reading comprehension test performance? In this chapter, the relationship between cognitive strategy use and reading comprehension performance will be discussed. There were two key inferential steps in the correlational analysis that was conducted. The first was to find out whether there was a relationship between cognitive strategy use and reading test performance. To carry out this step, the *p*-value was set to be 0.05 (i.e. 5 per cent was attributed to the limit of the probability of erroneously rejecting the null hypothesis that stated there was no relationship). If the *p*-value from the data was found to be less than 0.05 (e.g. 0.02), the null hypothesis could be rejected, thereby accepting that a relationship between the two variables existed. The rejection of the null hypothesis is known as *statistical significance*. In Phakiti's study, the *p*-value was found to be less than 0.05. It would be premature to conclude at this stage that a significant correlation has been found. This is because correlations depend on sample sizes and the

number of behavioural observations that constitute the numerical data to be used. The second step in the analysis is an evaluation stage in which the researcher asks whether the detected correlation coefficient is *useful* (e.g. Can it inform teaching practice? Can it confirm a theory?).

The second key step concerns the *practical significance of the study*. Researchers typically examine the effect size of a correlation coefficient. In a Pearson correctional analysis (Cohen 1988), Pearson *r* values of 0.10 (small), 0.30 (medium) and 0.50 (large) indicate the extent to which a detected effect size is practical (i.e. the larger the value, the higher the practical significance of the study). It is important not to confuse statistical significance with practical significance. Sometimes it is useful to know that there is no statistical significance since it can confirm a theory that states that two variables are not related. That is, when researchers cannot reject the null hypothesis (i.e. when $p > 0.05$), it does not always mean that the finding is unuseful. A non-statistical finding may confirm a theory. For example, if there is no statistical significance between a learning style and success in language learning, it may be useful to confirm a theory of learning that claims that while learning styles vary from individual to individual, they are not a major contributor to learning success.

Prior to the first key step, Phakiti examined the distribution of the data (by examining descriptive statistics and graphical displays to investigate whether there were two modes in the data distribution, which could affect the *r* coefficient). A *scatterplot* was used which matched the score of one variable with the score of the other variable for all participants, to check whether the two variables (i.e. cognitive strategy use and test performance) were related. Based on the examination of the scatterplot, it was observed that the relationship was linear and positive. After these steps had been carried out, the Pearson correlation was calculated ($p < 0.05$) and it was found that this indicated a positive relationship between cognitive strategy use and test performance. The correlation coefficient between cognitive strategy use and test performance was found to be 0.39. Researchers also examine the value of r^2 (i.e. r × r) to interpret the strength of the relationship between two variables. r^2 is known as *shared variance* and indicates the degree of overlap between the sets of observations (i.e. the strength of the relationship). A shared variance can also be treated as an *effect size* of a correlation.

In this study, the shared variance (r^2) between cognitive strategy use and reading comprehension performance was 0.15 (0.39 × 0.39); that is, there was 15 per cent shared variance. In this study, it was argued that this degree was reasonable because there were other factors influencing language test performance, including communicative language ability, test-taker characteristics, test-method facets and random error as outlined by Bachman and Palmer (1996). It was also argued that although this correlation coefficient appeared to be weak, it was an important one because if cognitive use could contribute to 15 per cent of the language test performance, we

knew that learners who successfully used this strategy would be at an advantage. In summary, Pearson's correlation allowed Phakiti to evaluate to what extent the two variables were related and whether the detected relationship was meaningful and interpretable.

Resources for further reading

Bachman, L 2004, *Statistical Analyses for Language Assessment*, Cambridge University Press, Cambridge.

This book presents both conceptual and statistical procedures useful for general applied linguistics research. It discusses the basics of quantitative data analysis and statistical tests.

Dörnyei, Z 2007, *Research Methods in Applied Linguistics: Quantitative, Qualitative, and Mixed Methodologies*, Oxford University Press, Oxford.

This book treats quantitative research comprehensively from the design stage through to the analysis and reporting stages.

Larson-Hall, J 2010, *A Guide to Doing Statistics in Second Language Research using SPSS*, Routledge, New York, NY.

This book is comprehensive in the treatment of statistics in second language research. It also provides procedures of how to perform statistical analyses in SPSS.

Lowie, W & Seton, B 2013, *Essential Statistics for Applied Linguistics*, Palgrave Macmillan, Hampshire.

This book explains both descriptive and inferential statistics in applied linguistics. It presents how to use SPSS for common statistical tests with examples of how a particular analysis can be done.

Mackey, A & Gass, S 2005, *Second Language Research*. Lawrence Erlbaum Associates, Mahwah, NJ.

This book provides an accessible account of the procedures of quantitative data analysis. Chapter 9 provides a discussion of issues involved in quantitative data analysis and procedures for carrying out statistical analysis.

Roever, C & Phakiti, A in press, 2016, *Quantitative Methods for Second Language Research: A Problem-solving Approach*, Routledge, London.

This book introduces quantitative approaches to data analysis in applied linguistics with particular emphasis on second language learning and assessment research.

Woodrow, L 2014, *Writing about Quantitative Research in Applied Linguistics*, Palgrave Macmillan, London.

This book is an accessible reference text which aims to help people write about quantitative research in applied linguistics. It explains different types of statistical analyses with annotated examples drawn from published and unpublished sources.

References

Altman, DG 1991, *Practical Statistics for Medical Research*, Chapman and Hall, London.

American Psychological Association (APA) 2010, *Publication Manual of the American Psychological Association*, 6th edn, American Psychological Association, Washington, DC.

Bachman, LF & Palmer, AS 1996, *Language Testing in Practice*, Oxford University Press, Oxford.

Brown, JD 1991, 'Statistics as a foreign language – Part1: What to look for in reading statistical language studies', *TESOL Quarterly*, vol. 25, no. 4, pp. 569–586.

——— 1992, 'Statistics as a foreign language – Part2: More things to consider in reading statistical language studies', *TESOL Quarterly*, vol. 26, no. 4, pp. 629–664.

——— 1996, *Testing in Language Programs*, Prentice Hall, Upper Saddle River, NJ.

Carr, N 2008, 'Using Microsoft Excel to calculate descriptive statistics and create graphs', *Language Assessment Quarterly*, vol. 5, no. 1, pp. 43–62.

Cohen J 1960, 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46.

Cohen, J 1988, *Statistical Power Analysis for the Behavioral Sciences*, Sage, Newbury Park, CA.

Dörnyei, Z 2007, *Research Methods in Applied Linguistics: Quantitative, Qualitative, and Mixed Methodologies*. Oxford University Press, Oxford.

Fleiss, JL 1981, *Statistical Methods for Rates and Proportions*, 2nd edn, John Wiley, New York, NY.

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

Guba, EG & Lincoln, YS 2005, 'Paradigmatic controversies, contradictions, and emerging confluences', in NK Denzin & YS Lincoln (eds), *The Sage Handbook of Qualitative Research*, 3rd edn, Sage, Thousand Oaks, CA, pp. 191–215.

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

Larson-Hall, J 2010, *A Guide to Doing Statistics in Second Language Research using SPSS*, Routledge, New York, NY.

Lazaraton, A 2005, 'Quantitative research methods', in E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 209–224.

Lightbown, PM & Spada, N 2013, *How Languages are Learned*, 4th edn, Oxford University Press, Oxford.

Loewen, S & Gass, S 2009, 'The use of statistics in L2 acquisition research', *Language Teaching*, vol. 42, no. 2, pp. 181–196.

Macaro, E 2010, *Continuum Companion to Second Language Acquisition*, Continuum, London.

Mackey, A & Gass, S 2005, *Second Language Research*. Lawrence Erlbaum Associates, Mahwah, NJ.

Ockey, GJ 2014, 'Exploratory factor analysis and structural equation modeling', in AJ Kunnan (ed.), *The Companion to Language Assessment*, John Wiley & Sons, London, pp. 1–21.

Ortega, L 2009, *Understanding Second Language Acquisition*, Hodder, London.
Phakiti, A 2003, 'A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance', *Language Testing*, vol. 20, no. 1, pp. 26–56.
——— 2007, *Strategic Competence and EFL Reading Test Performance*, Peter Lang, Frankfurt am Main.
——— 2014, *Experimental Research Methods in Language Learning*, Bloomsbury, London.
Plonsky, L 2013, 'Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research', *Studies in Second Language Acquisition*, vol. 35, no. 4, pp. 655–687.
——— 2014, 'Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform', *The Modern Language Journal*, vol. 98, no. 1, pp. 450–470.
Plonsky, L & Gass, S 2011, 'Quantitative research methods, study quality, and outcomes: The case of interaction research', *Language Learning*, vol. 61, no. 2, pp. 325–366.
Purpura, P 2011, 'Quantitative research methods in assessment and teaching', in E Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning*, vol 2. Routledge, New York, NY, pp. 731–751.