CHAPTER 4

Research Variables, Validity, and Reliability

This chapter focuses on the concepts necessary for understanding how to design a study in second language research. We begin with an outline of variables and scales, and follow with descriptions of specific types of validity and reliability. We also discuss sampling, representativeness and generalizability, and the collection of biographical data.

4.1. INTRODUCTION

In chapter 1, we introduced the concepts of research questions and research hypotheses. Research questions can take a range of forms. One example of a specific and answerable research question might be, "What is the effect of form-focused instruction on the acquisition of English relative clauses by French- and Japanese-speaking learners of English?" Because of differences between Japanese and English and similarities between French and English, we might hypothesize as follows: "French-speaking learners of English will perform better following form-focused instruction than will Japanese-speaking learners of English." Assuming that the research question is clearly phrased, answerable, and motivated by the literature, we can move on to the research hypotheses.

4.2. HYPOTHESES

A hypothesis is a type of prediction found in many experimental studies; it is a statement about what we expect to happen in a study. In research reports there are generally two types of hypotheses: research hypotheses and null hypotheses. The null hypothesis (often written as H_0) is a neutral statement used as a basis for testing. The null hypothesis states that there is no re-

lationship between items under investigation. The statistical task is to reject the null hypothesis and to show that there is a relationship between X and Y. Given our hypothesis above that French-speaking learners of English would perform better following form-focused instruction than would Japanese-speaking learners of English, the null hypothesis would be:

> There will be no difference between the performance of the French group and the Japanese group on a posttest.

We could then statistically test the differences in performance between these groups on a posttest following instruction to determine if any differences found were due to chance or due to treatment. We return to hypotheses and statistics in chapter 9.

When, based on previous research reports in the literature, we expect a particular outcome, we can form research hypotheses. There are two ways that we can do this. The first is to predict that there will be a difference between two groups, although we do not have sufficient information to predict the direction of the difference. For example, we might have a research hypothesis that states simply that the two groups will be different, such as:

> There will be a difference between the performance of the French-speaking group and the Japanese-speaking group on a posttest.

This is known as a nondirectional or two-way hypothesis.

On the other hand, we may have enough information to predict a difference in one direction or another. This is called a directional or one-way hypothesis. To continue our example, we might believe (based on the closer linguistic relationship between English and French than between English and Japanese) that the French-speaking group will perform better than the Japanese-speaking group. We would then formulate our hypothesis as follows:

> The French-speaking group will perform better on a posttest than the Japanese-speaking group.

4.3. VARIABLE TYPES

In order to carry out any sort of measurement, we need to think about variables; that is, characteristics that vary from person to person, text to text, or object to object. Simply put, variables are features or qualities that change.

TABLE 4.1
Variable Types

Research Question	*Independent Variable*	*Dependent Variable*
Does feedback type affect subsequent performance?	Feedback type	Performance measure
Can elements of child-directed speech aid in learning morphology?	Child-directed speech	Measure of morphological acquisition
Does length of residence affect identification of word-final consonants?	Length of residence	Measure of success in identifying word-final consonants
Is there a relationship between learners' noticing of recasts and L2 development?	Noticing of recasts	L2 development measure

For example, we might want to think about the effects of a particular pedagogical treatment on different groups of people (e.g., Spanish speakers learning English versus Japanese speakers learning English). Native language background, then, is a variable. What we are ultimately doing in experimental research is exploring whether there are relationships between variables and a constant:

Example:

We want to examine the effects of different types of instruction on a group of foreign language learners. We take as our object of investigation students enrolled in first semester foreign language classes of Spanish. We use two equivalent classes of first semester Spanish (selecting classes is an issue that we discuss in chap. 5). We have teachers provide one group with explicit grammar instruction. Another group receives no grammar instruction, but receives a significant amount of input on the specific linguistic structure in question.

The variable under investigation: Type of instruction.

What is being held constant: Class level (first semester Spanish, native language background of participants).

4.3.1. Independent and Dependent Variables

There are two main variable types: independent and dependent. The independent variable is the one that we believe may "cause" the results; the dependent variable is the one we measure to see the effects the independent variable has on it. Let us consider the examples in Table 4.1.

In each of the examples in Table 4.1, the independent variable is manipulated to determine its effect on the dependent variable. To elaborate on one of these, let us consider the third example: Does length of residence affect identification of word-final consonants? Let us assume that we have a well-motivated reason to believe that learners are able to recognize word-final consonants based on the amount of input to which they have been exposed. Assuming that one can operationalize (see sec. 4.4.) amount of input (possibly as length of residence in a foreign country or amount of classroom exposure), we would then divide learners into groups depending on their exposure to the target language and see if there is a difference in the degree of correct identification of word-final consonants between or among the groups. The dependent variable would be expressed in terms of the number or percentage of word-final consonants correctly identified, and we would determine whether learners with longer or greater exposure (independent variable) had higher scores (dependent variable).

It is clear that the variables in Table 4.1 also differ in another way—some can be directly manipulated by the researcher (e.g., feedback types), whereas some already exist (e.g., amount of input). With those that exist, the researcher needs to find the right way of selecting the appropriate forum for investigating the effects. With those that can be manipulated, the task of the researcher is to determine how to manipulate the variable appropriately. For example, in the case of feedback types, one could select three different teachers who naturally employ different feedback types and use their classrooms for investigation. Alternatively, one could train teachers to use different feedback types.

4.3.2. Moderator Variables

Moderator variables are characteristics of individuals or of treatment variables that may result in an interaction between an independent variable and other variables. Let us assume again a study on the effect of length of residence on the recognition of word-final consonants. Let us further assume that we have a theoretical rationale for believing that

length of residence might differentially affect recognition depending on gender. Gender might then be considered to be a moderator variable. In other words, a moderator variable is a type of independent variable that may not be the main focus of the study, but may modify the relationship between the independent variable and the dependent variable. Of course, moderator variables can "sneak" into a study without the researcher realizing that they may be important. One could imagine that in the hypothetical study mentioned earlier, it might not occur to the researcher that gender would be a factor. We have to be cognizant, therefore, of the fact that there may be variables that interfere with the actual results we are seeking. These are known as intervening variables.

4.3.3. Intervening Variables

Intervening variables are similar to moderator variables, but they are not included in an original study either because the researcher has not considered the possibility of their effect or because they cannot be identified in a precise way. For instance, consider a study that measures the effect of pedagogical treatment (independent variable) on learners' overall language proficiency (dependent variable, as measured by TOEFL scores). A variable that cannot be measured or understood easily might be the individuals' test-taking abilities. In other words, the results may be due to test-taking abilities rather than to the treatment. Because this variable was not controlled for, it is an intervening variable that could complicate the interpretation of the results.

4.3.4. Control Variables

When conducting research, one ideally wants to study simply the effects of the independent variable on a dependent variable. For example, consider the impact of feedback type on a performance measure. Variables that might interfere with the findings include the possibility that learners with different levels of proficiency respond differently to different types of feedback. Another possibility is that different students, depending on their prior language learning experiences, respond differently to different types of feedback. Whenever possible, researchers need to identify these possible factors and control for them in some way, although it should be recognized that identifying and controlling for all variables in L2 research may be difficult.

One way to determine if gender or possibly native language background (as a way of operationalizing language learning experiences) might have an

effect is to balance these variables by having an equal number of men versus women or an equal number of Korean versus Japanese versus Spanish speakers (or whatever languages one is dealing with). These then become moderator variables (see earlier discussion). Another way to control for possibly interfering, or confounding, variables is to eliminate the variable completely (i.e., to keep it constant). In our hypothetical example, our study might include only men or only women, or only Korean or Japanese or Spanish speakers. Gender and native language then become control variables. This latter solution, of course, limits the degree of generalizability of one's study (see sec. 4.6.7 on external validity).

4.4. OPERATIONALIZATION

In many instances in second language research it is difficult to measure variables directly so researchers provide working definitions of variables, known as operationalizations. An operational definition allows researchers to operate, or work, with the variables. Operationalizations allow measurement. To return to the earlier example, we said that we need to operationalize "amount of input," because this term, as stated, is vague. Although it might be difficult to come up with a uniform concept of amount of input, it is possible to think of examples in which groups vary along some parameter that seems close to the amount of input. For example, classroom learners could be classified based on how many years of exposure they have had to the target language, for example 1 versus 2 versus 3 years. Hence, "amount of input" could be operationalized as years of exposure in this case. In a more natural setting, the operationalization of "amount of input" could be the number of years spent in the target language environment. Once a variable has been operationalized in a manner such as this, it is possible to use it in measurements.

4.5. MEASURING VARIABLES: SCALES OF MEASUREMENT

We have discussed variable types, but there is another way that we can think of differences in variables: What scales are going to be used to describe and analyze the data from the different variables? In this section, we provide a brief introduction to different scales. Chapter 8 on data coding deals with the topic in greater detail.

The three most commonly used scales are nominal, ordinal, and interval. Ratio scales, a type of interval scale, are not included here because they

are not used as frequently in the type of research that is carried out in second language studies.[1] Nominal scales are used for attributes or categories and allow researchers to categorize variables into two or more groups. With nominal scales, different categories can be assigned numerical values. For example, in a study of gender, (1) may be assigned to male and (2) to female. The numbers indicate only category membership; there is no indication of order or magnitude of differences. Consequently, in a nominal scale the concept of average does not apply.

An ordinal scale is one in which ordering is implied. For example, student test scores are often ordered from best to worst or worst to best, with the result that there is a 1st-ranked student, a 2nd-ranked student, a 10th-ranked student, and so forth. Although the scores are ordered, there is no implication of an equal distance between each rank order. Thus, the difference between Students 1 and 2 may not be the same as the difference between Students 2 and 3. It is also often the case that researchers need to give holistic judgments to student work. This might be the case, for example, with second language writing scores. If we gave writing scores on a scale from 1 to 100, we might not be able to say that someone who received an 80 is twice as good a writer as someone who received a 40 without having precise information about what 40 and 80 meant on the scale. An ordinal scale might be useful in ordering students for placement into a writing program, but we cannot make judgments about exactly how much better one student is than another.

An interval scale represents the order of a variable's values, but unlike an ordinal scale it also reflects the interval or distance between points in the ranking. If a test represents an interval scale, then one can assume that the distance between a score of 70 and 80 is the same as the distance between 80 and 90. Thus, we could say, for example, that someone who received a score of 10 on a vocabulary test knew twice as many of the words that were tested as did someone who received a 5. As this example shows, an interval scale implies measurable units, such as number of correct answers, years of residence in the target language country, or age (as opposed to being a good writer).

4.6. VALIDITY

After spending a great deal of time and effort designing a study, we want to make sure that the results of our study are valid. That is, we want them to reflect what we believe they reflect and that they are meaningful in the

[1] Ratio scales have a true zero point where zero represents the absence of the category.

sense that they have significance not only to the population that was tested, but, at least for most experimental research, to a broader, relevant population. There are many types of validity, including content, face, construct, criterion-related, and predictive validity. We deal with each of these in turn before turning to internal and external validity, which are the most common areas of concern.

4.6.1. Content Validity

Content validity refers to the representativeness of our measurement regarding the phenomenon about which we want information. If we are interested in the acquisition of relative clauses in general and plan to present learners with an acceptability judgment task, we need to make sure that all relative clause types are included. For example, if our test consists only of sentences such as "The boy who is running is my friend," we do not have content validity because we have not included other relative clause types such as "The dog that the boy loves is beautiful." In the first sentence the relative pronoun *who* is the subject of its clause, whereas in the second sentence the relative pronoun *that* is the object. Thus, our testing instrument is not sensitive to the full range of relative clause types, and we can say that it lacks content validity.

4.6.2. Face Validity

Face validity is closely related to the notion of content validity and refers to the familiarity of our instrument and how easy it is to convince others that there is content validity to it. If, for example, learners are presented with reasoning tasks to carry out in an experiment and are already familiar with these sorts of tasks because they have carried them out in their classrooms, we can say that the task has face validity for the learners. Face validity thus hinges on the participants' perceptions of the research treatments and tests. 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.

4.6.3. Construct Validity

This is perhaps the most complex of the validity types discussed so far. Construct validity is an essential topic in second language acquisition research precisely because many of the variables investigated are not easily or di-

rectly defined. In second language research, variables such as language proficiency, aptitude, exposure to input, and linguistic representations are of interest. However, these constructs are not directly measurable in the way that height, weight, or age are. In research, construct validity refers to the degree to which the research adequately captures the construct of interest. Construct validity can be enhanced when multiple estimates of a construct are used. For example, in the hypothetical study discussed earlier that was seeking to link exposure to input with accuracy in identifying final consonants, the construct validity of the measurement of "amount of input" might be enhanced if multiple factors—such as length of residence, language instruction, and the language used in the participants' formal education—were considered together.

4.6.4. Criterion-Related Validity

Criterion-related validity refers to the extent to which tests used in a research study are comparable to other well-established tests of the construct in question. For example, many language programs attempt to measure global proficiency either for placement into their own program or to determine the extent to which a student might meet a particular language requirement. For the sake of convenience, these programs often develop their own internal tests, but there may be little external evidence that these tests are measuring what the programs assume they are measuring. One could measure the performance of a group of students on the local test and a well-established test (e.g., TOEFL in the case of English, or in the case of other languages, another recognized standard test). Should there be a good correlation (see chap. 9 for a discussion of correlations in statistics), one can then say that the local test has been demonstrated to have criterion-related validity.

4.6.5. Predictive Validity

Predictive validity deals with the use that one might eventually want to make of a particular measure. Does it predict performance on some other measure? Considering the earlier example of a local language test, if the test predicts performance on some other dimension (class grades), the test can be said to have predictive validity.

We now turn to the two main types of validity that are important in conducting research: internal validity and external validity.

4.6.6. Internal Validity

Internal validity refers to the extent to which the results of a study are a function of the factor that the researcher intends. In other words, to what extent are the differences that have been found for the dependent variable directly related to the independent variable? A researcher must control for (i.e., rule out) all other possible factors that could potentially account for the results. For example, if we wanted to observe reaction times to a set of grammatical and ungrammatical sentences, we might devise a computer program that presents sentences on a computer screen one at a time, with learners responding to the acceptability/unacceptability of each sentence by pressing a button on the computer. To make the task easier for the participants in the study, we could tape the letter *A* for "acceptable" over the letter *t* on the keyboard and tape the letter *U* for "unacceptable" over the *y* key on the keyboard. After we have completed the study, someone might ask us if we checked for handedness of the participants. In other words, could it be the case that for those who are left handed, the *A* key ("acceptable") might be faster not because it is faster to respond to acceptable as opposed to unacceptable sentences (part of our hypothesis), but because left hands on left-handed people react faster. Our results would then have been compromised. We would have to conclude that there was little internal validity.

It is important to think through a design carefully to eliminate or at least minimize threats to internal validity. There are many ways that internal validity can be compromised, some of the most common and important of which include participant characteristics, participant mortality (dropout rate), participant inattention and attitude, participant maturation, data collection (location and collector), and instrumentation and test effects.

4.6.6.1. Participant Characteristics

The example provided in the previous section concerning handedness is a participant characteristic. Clearly, not all elicitation techniques will require controlling for handedness. In other words, there may be elements of the research questions and/or elicitation technique that require a careful selection of one characteristic or another. Let us consider some relevant participant characteristics for second language research: language background, language learning experience, and proficiency level.

Language Background. In many studies, researchers want to compare one group of students with another group based on different treat-

ments. For example, let us assume that a study on the role of attention in second language learning compared groups of students in a foreign language class who were exposed to a language structure with and without devices to ensure that they paid attention to that structure. It would be important that each group of students be relatively homogeneous. Were they not homogeneous, one could not be sure about the source of the results. For instance, let's further assume that one group of students had a large number of participants who were familiar with a language closely related to the target language (either through exposure at home or in the classroom). We then could not distinguish between the effects of the treatment and the effects of the background knowledge of the participants.

Language Learning Experience. Participants come to a language learning situation with a wide range of past experiences. In some instances, these experiences may have importance for research. For example, many students in an ESL setting have had prior English instruction in their home countries, and this prior instruction may differ from one country to another. If we wanted to conduct a study in which we compared implicit versus explicit methods of instruction, we might find that a group that received explicit instruction outperformed a group that received implicit instruction. If the two groups also differed in terms of prior learning experiences, we would be left with two variables: learning experience and instruction type. We would not be able to distinguish between them. Are our results possibly due to the fact that explicit instruction yielded a better outcome because one group was more familiar with and thus more affected by that type of instruction? Or did one instruction type yield better results due to that type of instruction? It is the latter that we want our study to measure.

Proficiency Level. This is one of the most difficult areas to control for when conducting second language research. In foreign language environments, the issue is perhaps simpler than in second language environments, because in the former but not the latter there is limited exposure outside the classroom although here, too, there can be problems. In the area of foreign language research, there are some global proficiency measures such as the Oral Proficiency Interview (OPI) so that learners can be matched for proficiency. Another common measure is to use placement in class level (first year versus second year versus third year, etc.). In a foreign language environment, this is relatively "safe" because exposure is more or

less limited to what occurs in the classroom.[2] However, with second language learners, backgrounds and outside experiences are varied and there is typically unevenness in skill levels. For example, some potential participants in the same class level may have excellent oral skills but weak written skills, and vice versa. It is therefore important to consider how this may bear on the specific research questions of the study.

We have discussed some of the ways in which participant characteristics differ. It is also important to ensure that participants are matched on the feature that is being examined. For example, if one is conducting a study that investigates the perception and production of phonological categories, it may not be sufficient to assume that advanced students are better than intermediate students because the intermediate students may have spent more time in the country where the language is spoken than the advanced students and, consequently, their perception and production of target language sounds may be more advanced even if their command of other aspects of the language is not. One must also be wary of using global proficiency tests when the testing instrument relies on one skill or another. For example, a global language test that provides information on grammar and vocabulary may obscure differences in participants' listening abilities. If listening is a major part of gathering data (e.g., as in elicited imitation tasks, discussed in chap. 3), an additional measure of listening ability may be needed to make sure that difficulty with the instrument is not an issue causing problems with internal validity.

4.6.6.2. Participant Mortality

Some studies that are conducted in second language research are longitudinal in nature. That is, they seek to measure language development by sampling over time. As such, researchers may typically carry out immediate posttests and also one or more delayed posttests to determine the shorter- and longer-term effects of a treatment. In order to appropriately address research questions, it is best to ensure that all participants are present for all sessions. However, in many classroom research settings, it is inevitable that not all participants will be present at all times. A researcher must determine how to deal with this situation, and there a number of factors

[2]This is, of course, an oversimplification, because classroom learners will vary greatly in terms of the amount of time they spend out of class reading the foreign language or in the language laboratory.

that one might want to consider. For example, if a researcher has 50 participants and one of them has to be eliminated, the loss is probably not significant. If, on the other hand, participant numbers are balanced across groups, the loss of a participant in one group may necessitate the elimination of a matched participant in another group. Some possible scenarios follow.

Scenario 1: Participant Missing From One Treatment Session

Purpose of study:	Measuring the effect of quantities of input across groups.
Number of participants:	25 per group.
Method:	Differing amounts of input per lesson; five lessons over a 2-week period.
Posttests:	One posttest.
Situation:	One student in one group misses one class period.
Issue:	Should the posttest data for that student be included in the final data pool?
Response:	It might depend on how the groups vary in terms of input. Given that this study is measuring quantities of input and that one group may vary from others by only small differences in the amount of input, the inclusion of someone who missed one class session might make him or her more like someone in another group. Thus, data from this learner should probably be eliminated.

Scenario 2: Participant Missing From One Posttest

Purpose of study:	Determining the long-term effects of attention.
Number of participants:	Two groups of 10 each.
Method:	Computer-based input varying attention conditions.
Posttests:	Five posttests given at 1-month intervals.
Situation:	One student in one group misses one posttest.
Issue:	Should the data for that student be included in the final data pool?
Response:	Given that there are five post-tests, one could make a decision in advance that a student must

participate in at least the first and the last of the posttests and two of the remaining three. This would allow some flexibility in keeping as many participants in the data pool as possible while still providing the researcher with information from four data points following the treatment.

Scenario 3: Participant Missing From Part of Posttest

Purpose of study:	Determining the long-term effects of attention on syntax versus vocabulary.
Number of participants:	Two groups of 10 each.
Method:	Computer-based input varying attention conditions.
Posttests:	One posttest for syntax and one for vocabulary.
Situation:	One student in one group misses one posttest (either syntax or vocabulary).
Issue:	Should the data for that student be included in the final data pool?
Response:	Given that there are two separate posttests and assuming that data are being aggregated rather than each student's performance on syntax being compared to his or her performance on vocabulary, one could maintain the data in the pool. This would mean that data for the statistical tests would include 10 syntax scores versus 9 vocabulary scores (or vice versa). If, on the other hand, one wanted to do a comparison of each person's data on the two tasks, then the individual who missed one would not be able to be kept in the final data pool.

We have presented three sample scenarios showing what some of the considerations might be in determining what to do about participant mortality. Each situation will, undoubtedly, be different and will require a different (and justifiable) solution. The point to remember is the importance of carefully thinking through the various possibilities given the design of the experiment or longitudinal sessions, and making a principled decision as to how to solve the problem in the event of participant absences. These decisions should not be made ad hoc; when possible, they

should be made in advance of data collection. They should also be fully reported in the research report.

4.6.6.3. Participant Inattention and Attitude

When we collect data from participants, we usually make the assumption that they are giving us their "best effort." In other words, we rely on the notion that the language data we are collecting are uncontaminated by the experiment itself. This may not always be true. One factor that might affect participant behavior is what is known as the Hawthorne effect, which refers to the positive impact that may occur simply because participants know that they are part of an experiment and are, therefore, "different" from others. Participants may also try to please the researcher by giving the answers or responses they think are expected. This is known as the halo effect. Hawthorne and halo effects are also discussed in chapters 6 and 7 in relation to experimental designs and qualitative research.

Participating in a study also has potential negative effects. For example, researchers might want to consider factors such as fatigue and boredom when asking participants to perform tasks. This was mentioned in chapter 3 in the discussion of the number of sentences to use in an acceptability judgment test. Whatever method is being used to gather data, one needs to think of the exhaustion and boredom factor. How much time can one reasonably ask a participant to perform without losing confidence in the results, especially if it is a repetitive and demanding task such as judging sentences? There is no magic answer; we must weigh the need to gather sufficient data against these factors. As discussed earlier, presenting tasks or items in different orders can serve to balance these effects.

A second factor is general inattentiveness, whether from the outset of the experiment or as a result of the experiment. In a study by Gass (1994) that involved giving participants the same task after a 1-week interval, the author noted that some participants provided diametrically opposed responses at the two time periods. In a stimulated recall after the second session, one of the participants stated that his results from the two sessions differed because his mind was wandering given that he had two academic tests that week. Clearly, one does not always know whether this is an issue, but one needs to be aware of this as a possible way of explaining what may appear to be aberrant or divergent results. In general, if time and resources permit, it is helpful to do a stimulated recall with participants (possibly using the test measure as a stimulus) or a postexperiment interview or exit

questionnaire to ascertain if there might be extra-experimental factors that impacted learner responses or behaviors. Gathering such data from even a subset of participants can help in interpreting results.

4.6.6.4. Participant Maturation

Maturation is most relevant in longitudinal studies and particularly in those involving children. For example, a study that spans a year or longer will inevitably include participants who change in one way or another in addition to changes in language development. Adults may not change dramatically in a 1-year period, but children certainly do. Moreover, people who were comparable at the outset of the study may change in different ways due to different experiences over time. Thus, one must find a way to balance regular maturational factors against the requirements of the study. When maturation is a consideration, a control group not subjected to the treatment or intervention is appropriate wherever possible. The inclusion of a control group provides one way to test whether any changes occurred because of the experimental treatment or because of maturation.

4.6.6.5. Data Collection: Location and Collector

Not all research studies will be affected by the location of data collection, but some might. Some obvious concerns relate to the physical environment; for example, the environment for two groups given the same test might influence the results if one group is in a noisy or uncomfortable setting and the other is not. A perhaps less obvious effect of setting might occur in a study in which a researcher is trying to gather information from immigrant parents (perhaps through an oral interview) about their attitudes concerning their desires for their children to learn the target language. Informal interviews in their home might yield results that differ from those obtained in a formal school setting, where teachers in proximity could influence what the parents think they should say.

Another factor in some types of research relates to the person doing the data collection. Given the scenario mentioned earlier concerning families being surveyed about their attitudes toward their children's learning of the target language, one could imagine different results depending on whether or not the interviewer is a member of the native culture or speaks the native language.

4.6.6.6. Instrumentation and Test Effects

The test instrument is quite clearly an important part of many research studies. In this section we discuss three factors that may affect internal validity: equivalence between pre- and posttests, giving the goal of the study away, and test instructions and questions.

Equivalence Between Pre- and Posttests. One serious design issue relates to the comparability of tests. A difficult pretest with an easier post-test will make it more likely for improvement to be apparent after a treatment. The opposite scenario will make it more likely for no improvement to be apparent following a treatment. There are a number of ways to address comparability of tests. For example, when testing grammatical improvement following a treatment, one can keep the grammatical structure the same and change the lexical items. Doing this, however, requires ensuring comparable vocabulary difficulty. For example, the sentence *The dog ate the chair* does not involve the same vocabulary difficulty level as *The deer consumed the rhododendron*. One way to address this issue might involve consulting a word frequency index (e.g., Brown Corpus, Academic English, Academic Word List; see Francis & Kucera, 1982; Thorndike & Lorge, 1944) that lists words of the same frequency—that is, words that appear approximately the same number of times in a corpus of the same size and type.

Another way to ensure comparability is to establish a fixed group of sentences for all tests. If a set of 30 sentences were established, Participant A could have a random set of 15 of those on the pretest and the remaining 15 on the posttest. Participant B could also have a random set of 15 on the pretest and the remaining is on the posttest, but the two participants would in all likelihood not have the same sets of 15. This is quite easy to do on a computer, but it could be done without a computer as well, counterbalancing the test by giving half of a group one set of sentences on the pretest and the other set on the posttest and giving the sets of sentences to the other half of the group in the reverse order. This technique may also eliminate the possible practice effects or participant inattentiveness that might arise if learners were tested on the same set of sentences twice.

Another example of the importance of test comparability can be seen in conducting second language writing studies. Researchers need to be mindful of the need to choose appropriate topics about which to write. A pretest that is based on a compare and contrast essay might be quite different in structure and vocabulary than a posttest essay based on a topic of persuasion. It would not be meaningful to compare the two essays.

Giving the Goal of the Study Away. One of the problems in doing second language research is that one sometimes does not want participants to know the precise nature of the language area or behavior that is being tested. We might want to conceal the precise nature of the study because we want responses that reflect natural behavior rather than what participants think they should say or do (see chap. 2 for a discussion of consent forms and how to strike a balance between not being deceptive and yet not revealing precisely what the study's focus is). This becomes particularly problematic when using a pretest because the pretest may in and of itself alert participants to the study's objective. One way of avoiding this problem is by conducting the pretest a few weeks before the study, the idea being that participants will not associate the pretest with the study itself. The disadvantage of this solution is that in the time interval between the pretest and the actual treatment and posttest, participants' knowledge may change, making the results unreliable. A modification of this is to have a shorter time differential, but that, of course, weakens the original issue—that of not revealing the topic of the study. A second solution, particularly in the case of assessment of discrete language knowledge, is to ensure that the grammatical/lexical point in question is embedded in a much larger test, thereby reducing the likelihood of participants figuring out the scope of the test. If the participants do not guess the topic from the pretest, the study instruments are more likely to produce a valid characterization of their L2 knowledge.

Instructions/Questions. In addition to guarding against the previously discussed threats to internal validity, one must make sure that the instructions are clear and appropriate to the developmental level of the participants in the study. We cannot rely on responses to questions when it is not clear whether the instructions have been adequately understood. For example, on a university application (filled out by native as well as non-native speakers of English) are the following questions "1. Have you ever been expelled, suspended, disciplined, or placed on probation by any secondary school or college you have attended because of (a) academic dishonesty, (b) financial impropriety, or (c) an offense that harmed or had the potential to harm others?" and "2. Have your ever been convicted of a criminal offense (including in juvenile court) other than a minor traffic violation or are there criminal charges pending against you at this time?" This is followed by "If circumstances arise in the future (until the time you begin attending classes) that make your answers to the above questions inaccurate, misleading, or incomplete, you must provide the Office of Admissions with up-

dated information." As can be seen, the language is overly sophisticated for those whose English language abilities are not nativelike, and the content (e.g., juvenile court) is inappropriate for a wide range of students who may come from countries with different legal systems. In second language research, the instructions and questions should be appropriate to the level of linguistic and cultural knowledge of those who are taking the test or filling out a questionnaire.

This section has dealt with threats to internal validity. A summary of ways in which such threats can be minimized includes:

- Consider *participant characteristics* that may be relevant to the research questions and elicitation techniques, including but not limited to:

  Language background.
  Past language learning experiences.
  Proficiency level.
  Specific features and/or skills being examined.

- Consider the issue of *participant mortality*. Make decisions about it before carrying out your research, and justify your solution with respect to:

  Research design.
  Research questions.
  How significant the loss of data would be.
  (Then be sure to report on this in your research article.)

- Be aware of the possibility that the experimentation itself may affect the results through:

  Hawthorne and halo effects.
  Fatigue and boredom of participants.
  Practice effects of the test material.

- Get the participants' perspectives after the experiment to ascertain if extra-experimental factors may have impacted their behavior.
- Use a control group to balance *maturational factors* against any long-term requirements of the study.
- Consider how the participants' performance might be affected by:

  Physical environment of the study.
  Characteristics of the researcher.

- Ensure the comparability of pre- and posttests.

- Don't give away the goals of the study.
- Make sure that the instructions are clear and appropriate to the developmental level of the participants.

In the next section, we turn to another type of validity, that known as external validity.

4.6.7. External Validity

All research is conducted within a particular setting and using a specific set of characteristics (e.g., second year L1 English learners of French at X university). However, most quantitative research is concerned with broader implications that go beyond the confines of the research setting and participants. The participants chosen for any study form a research population. With external validity, we are concerned with the generalizability of our findings, or in other words, the extent to which the findings of the study are relevant not only to the research population, but also to the wider population of language learners. It is important to remember that a prerequisite of external validity is internal validity. If a study is not conducted with careful attention to internal validity, it clearly does not make sense to try to generalize the findings to a larger population.

4.6.7.1. Sampling[3]

The basis of generalizability is the particular sample selected. We want our particular group of participants to be drawn randomly from the population to which we hope to generalize. Thus, in considering generalizability, we need to consider the representativeness of the sample. What this means is that each individual who could be selected for a study has the same chance of being selected as does any other individual. To understand this, we introduce the concept of random sampling.

Random Sampling. Random sampling refers to the selection of participants from the general population that the sample will represent. In most second language studies, the population is the group of all language learners, perhaps in a particular context. Quite clearly, second language researchers do not have access to the entire population (e.g., all learners of

[3]As mentioned earlier, the sampling procedures discussed in this section relate primarily to quantitative studies. Qualitative research is discussed in chapter 6.

Spanish at U.S. universities), so they have to select an accessible sample that is representative of the entire population.

There are two common types of random sampling: simple random (e.g., putting all names in a hat and drawing from that pool) and stratified random sampling (e.g., random sampling based on categories). Simple random sampling is generally believed to be the best way to obtain a sample that is representative of the population, especially as the sample size gets larger. The key to simple random sampling is ensuring that each and every member of a population has an equal and independent chance of being selected for the research. However, simple random sampling is not used when researchers wish to ensure the representative presence of particular subgroups of the population under study (e.g., male versus female or particular language groups). In that case, stratified random sampling is used.

In stratified random sampling, the proportions of the subgroups in the population are first determined, and then participants are randomly selected from within each stratum according to the established proportions. Stratified random sampling provides precision in terms of the representativeness of the sample and allows preselected characteristics to be used as variables. In some types of second language research it might be necessary, for example, to balance the number of learners from particular L1 backgrounds in experimental groups. For other sorts of second language questions it might be important to include equal numbers of males and females in experimental groups, or to include learners who are roughly equivalent in terms of amount and type of prior instruction or length of residence in the country where the research is being conducted. As an example, assume that one is conducting a study on the acquisition of Arabic passives by speakers of English. Let's further assume that in Arabic language programs, there is a mixture of heritage speakers (those learners who have been exposed to Arabic prior to formal language study through family situations) and nonheritage speakers. Of the students who are available for the study, it turns out that 75% are heritage speakers, making it unlikely that the results will be generalizable to all learners of Arabic. To avoid this problem, the researcher could decide to obtain a sample containing 50% heritage learners and 50% nonheritage learners and randomly select accordingly. This would also make possible what might be an important comparison—that of heritage versus nonheritage learners.

There is yet another approach to sampling, called cluster random sampling. Cluster random sampling is the selection of groups (e.g., intact second language classes) rather than individuals as the objects of study. It is

more effective if larger numbers of clusters are involved. In larger-scale second language research, for example, it might be important to ensure that roughly equal numbers of morning and evening classes receive the same treatments; however, as with any method, the research question should always drive the sampling choice.

How does one obtain a random sample? As mentioned earlier, the principle that should guide selection is that each member of the population has an equal and independent chance of being selected. The purest way of obtaining a true random sample is to take all members of the possible sample, assign each a number, and then use a random number table (available from most statistics books) or a computer-generated random number table (for example, using Microsoft Excel). The following is a small random number table:

068273	241371
255989	213535
652974	357036
801813	313669
188238	987762
858182	324564
539567	010407
874905	076754
705832	752953
394208	866085
532487	980193
717734	499039
965606	256844
442732	809259
128056	843715
398907	972289
999451	782983
016511	525925
980529	329844
657643	501602
123905	385449
941465	573504
311991	088504
594989	631367
163091	221076

If, for example, you have an available population of 99 but you only want to use 35 individuals for your study, you could assign each member a number from 1 to 99 and then use a random number generator to select the first 35. If the first number generated is 77, the person who has been assigned 77 will be part of the data pool. Let's assume that the second number generated is 55. The person who has been assigned 55 will also be a member of the data pool. This continues until 35 numbers have been generated. Alternatively, using the random number table just presented, you could decide to use the last two digits (or the first two or the middle two) and select the first 35 numbers that fall between 01 and 99 until you have the 35 individuals that you need for your study. Starting from the left column and using the last two digits, you would select 73, 89, 74, 13, 38, 82, 67, and so on until you had 35 participants.

Nonrandom Sampling. Nonrandom sampling methods are also common in second language research. Common nonrandom methods include systematic, convenience, and purposive sampling. Systematic sampling is the choice of every *n*th individual in a population list (where the list should not be ordered systematically). For example, in organizing a new class where learners have seated themselves randomly in small groups (although one must be sure that the seating was truly random rather than in groups of friends/acquaintances), teachers often ask learners to count themselves off as As, Bs, and Cs, putting all the As into one group and so on. In a second language study, researchers could do the same for group assignments, although it would be important that the learners were seated randomly.

Convenience sampling is the selection of individuals who happen to be available for study. For instance, a researcher who wanted to compare the performance of two classes after using different review materials might select the two classes that require the review materials based on the curriculum. The obvious disadvantage to convenience sampling is that it is likely to be biased and should not be taken to be representative of the population. However, samples of convenience are quite common in second language research. For example, researchers may select a time and a place for a study, announce this to a pool of potential participants, and then use those who show up as participants. These learners will show up depending on their motivation to participate and the match between the timetable for the research and their own schedules and other commitments.

In a purposive sample, researchers knowingly select individuals based on their knowledge of the population and in order to elicit data in which they are interested. The sample may or may not be intended to be representa-

tive. For example, teachers may choose to compare two each of their top-, middle-, and lower-scoring students based on their results on a test, or based on how forthcoming these students are when answering questions about classroom processes. Likewise, a researcher may decide to pull out and present in-depth data on particular learners who did and did not develop as a result of some experimental treatment in order to illustrate the different pathways of learners in a study. Some consequences of nonrandom sampling are discussed later in this chapter.

4.6.7.2. Representativeness and Generalizability

If researchers want the results of a particular study to be generalizable, it is incumbent upon them to make an argument about the representativeness of the sample. Similarly, it is important to describe the setting. A study conducted in a university setting may not be generalizable to a private language school setting. It is often the case that to protect the anonymity of participants, one makes a statement such as the following about the location of the study: "Data were collected from 35 students enrolled in a second-year Japanese class at a large U.S. university." It is important to minimally include this information so that one can determine generalizability. Private language school students may be different from students at large universities, who may in turn be different from students at other types of institutions.

When choosing a sample, the goal is usually that the sample be of sufficient size to allow for generalization of results, at least for most nonqualitative sorts of research. It is generally accepted that larger samples mean a higher likelihood of only incidental differences between the sample and the population. To reflect this, many statistical tests contain built-in safeguards that help prevent researchers from drawing unwarranted conclusions.

Novice researchers often wonder how many learners are "enough" for each group or for their study overall.[4] In second language research,

[4]At a large university, a chemist, a physicist, and a statistician were meeting with their Provost in a conference room to explain the real-life applications of their disciplines. During the meeting, a fire broke out in a wastebasket. The physicist whipped out a calculator and began crunching numbers, explaining, "I'm calculating the amount of energy that must be removed in order to stop the combustion." The chemist thoughtfully examined the fire and jotted down some notes, explaining. "I'm figuring out which reagent can be added to the fire to prevent oxidation." The Provost seemed impressed at the speed of their reactions and exclaimed, "I had no idea that there could be such immediate real-world applications of your disciplines to a situation like this." Meanwhile, the statistician pulled out a book of matches and began to set all the other wastebaskets on fire. The shocked Provost demanded, "What are you doing? Are you crazy?" "No, not at all," replied the statistician, "It's just that we won't understand anything until we have a larger N!"

participant numbers vary enormously because of the wide range of different types of research conducted. These research types can range from an intensive experiment including several treatments, pretests, immediate posttests, and multiple delayed posttests, all entailing complex and finely grained linguistic analyses, to a large-scale second language testing study, in which simple numerical before and after scores may be utilized for hundreds of learners. In their text directed at educational research, Fraenkel and Wallen (2003) provided the following minimum sample numbers as a guideline: 100 for descriptive studies, 50 for correlational studies, and 15 to 30 per group in experimental studies depending on how tightly controlled they are. We must remember, however, that research in general education tends to have access to (and to utilize) larger pools than second language research. In second language studies, small groups are sometimes appropriate as long as the techniques for analysis take the numbers into account.

As we have said, a sample must be representative of the population in order for the results to be generalizable. If it is not representative, the findings have limited usefulness. If random sampling is not feasible, there are two possible solutions: First, thoroughly describe the sample studied so that others can judge to whom and in what circumstances the results may be meaningful. Second, as we also discussed in chapter 1, conduct replication studies (and encourage the same of others) wherever possible, using different groups of participants and different situations so that the results, if confirmed, may later be generalized.

4.6.7.3. Collecting Biodata Information

When reporting research, it is important to include sufficient information to allow the reader to determine the extent to which the results of your study are indeed generalizable to a new context. For this reason, the collection of biodata information is an integral part of one's database. The major consideration is how much information to collect and report with respect to the participants themselves. In general, it is recommended that the researcher include enough information for the study to be replicable (American Psychological Association, 2001) and for our purposes in this chapter enough information for readers to determine generalizability. However, the field of second language research lacks clear standards and expectations for the reporting of data, and instances of underreporting are frequent.

In reporting information about participants, the researcher must balance two concerns. The first is the privacy and anonymity of the partici-

Name _____ Research code _____

Gender: ____Male ____Female Age ____ First language(s) _____

E-mail address _____ Phone number _____

For how many years have you studied English? _____

How old were you when you started to study English? _____

_____

Where have you studied English? (tick as many as needed)	How long? (years)	Native English speaker? (yes/no)
_____ Kindergarten	_____	_____
_____ Elementary school	_____	_____
_____ Lower high school	_____	_____
_____ Upper High school	_____	_____
_____ Language schools	_____	_____
_____ Private Tutoring	_____	_____

What English classes are you studying in now? (class numbers and names)

_____

What English classes will you be taking next semester? (Class numbers and names)

_____

Are you studying English anywhere else now? Where? What are you studying (TOEFL, grammar)?

_____

_____

What was your score on the English test of the entrance exam? _____

Have you ever taken the TOEFL test? Yes____ No____ What was your score? _____

_____

How many hours per week do you spend using English outside class to …

Do homework	0	1–2	3–4	5–6
Prepare for quizzes and exams	0	1–2	3–4	5–6
Listen to language tapes	0	1–2	3–4	5–6
Read for fun	0	1–2	3–4	5–6
Listen to music	0	1–2	3–4	5–6
Watch TV, videos & movies	0	1–2	3–4	5–6
Talk to friends	0	1–2	3–4	5–6
Talk to tourists	0	1–2	3–4	5–6
Talk to family members	0	1–2	3–4	5–6

(continued on next page)

Have you ever been to an English-speaking country (UK, Canada, USA, Australia, etc.)?
Yes___ No___
If yes, how long were you there?_____ What did you do there?_____
Have ever been to a country where you spoke English to communicate (Japan, Malaysia, Vietnam, etc.)? Yes___ No ___
If yes, how long were you there? _____
Besides your first language and English, do you know any other languages? Yes___ No___
If yes, which languages? _____
How well do you know them? _____

FIG. 4.1. Sample biodata form.

pants; the second is the need to report sufficient data about the participants to allow future researchers to both evaluate and replicate the study. There are no strict rules or even guidelines about what information should be obtained in the second language field; because of this, exactly what and how much detail is obtained will depend on the research questions and will vary for individual researchers.

It is generally recommended that major demographic characteristics such as gender, age, and race/ethnicity be reported (American Psychological Association, 2001), as well as information relevant to the study itself (e.g., the participants' first languages, previous academic experience, and level of L2 proficiency). Additional information that might be important for a study on second language learning could include the frequency and context of L2 use outside the classroom, amount of travel or experience in countries where the L2 is spoken, learners' self-assessment of their knowledge of the target language, and the participants' familiarity with other languages. Additional information sometimes requested on biodata forms are facts that, although not appropriate for reporting, are necessary for carrying out the research, such as contact information and the association of the participant's name with a code number. *The Publication Manual of the American Psychological Association* also suggested that in reporting information about participants, selection and assignment to treatment groups also be included. The *Manual* further pointed out that "even when a characteristic is not an analytic variable, reporting it may give readers a more complete understanding of the sample and often proves useful in meta-analytic studies that incorporate the article's results" (American Psychological Association, 2001, p. 19).

A sample biodata form appears in Fig. 4.1. As can be seen from the form, depending on the data collection situation some of the questions might require explanations. Not all learners would automatically understand "first

language(s)," for example. Does it mean chronologically first? Does it mean *best* language? They might be more easily able to answer a question about which language they speak at home, or a more specific question about the first language learned and still spoken, or they might understand the term "mother tongue."

In devising forms for the collection of biographical data, it is important for researchers to balance their need for answers to the questions that could impact their study with requests for extra information that take time to elicit and explain. However, biographical information can be very important when selecting participants; for example, the form in Fig. 4.1 might elicit information about visits to English-speaking countries from even those learners who self-selected into a study on the basis of being beginners, but who then perform at a much higher level than the other learners in the study. This could be important in interpreting results. When selecting the precise questions it is necessary to consider how the data from the bodata form will be analyzed. For example, the form in Fig. 4.1 might be useful if one wants to categorize the amount of time spent using the L2 into four categories (none, little, moderate, a lot). A researcher might have used the number of hours on the form to ensure that her category of moderate was the same for all respondents (3–4 hours). However, if one is going to quantify these numbers across participants, these numbers are not easy to work with, particularly if one wants to combine categories into subcategories (e.g., listening). In other words, if a researcher were interested in listening, the categories "listen to language tapes," "listen to music" and "watch TV" might be combined. The difficulty in interpretation comes when trying to add the numbers. If someone responded 1–2, 3–4 and 3–4, the actual number of hours spent listening could be between 7 and 10, a range likely to be too great to be useful.

There may, however, be instances in which generalizability is not an issue. For example, if one is concerned about making curriculum changes or changes in the way assessment takes place in a particular language program, a research study may be conducted within the borders of that program. The results may turn out to be interesting enough to publish, but it should be understood that the results may or may not be applicable to other contexts and that it is only through empirical study in other contexts that one can determine the generalizability of the original findings.

In this section, we have pointed out that it is often difficult to ensure external validity but have shown ways to minimize threats to external validity. Following is a summary of ways in which one can deal with such threats:

- Random sampling.
- Stratified random selection.
- Systematic, convenience, and purposive sampling.
- Sufficient descriptive information about participants.
- Description of setting.
- Replication of study in a variety of settings.

4.7. RELIABILITY

Reliability in its simplest definition refers to consistency, often meaning instrument consistency. For example, one could ask whether an individual who takes a particular test would get a similar score on two administrations of the same test. If a person takes a written driving test and receives a high score, it would be expected that the individual would also receive a high score if he or she took the same written test again. We could then say the test is reliable. This differs from validity, which measures the extent to which the test is an indication of what it purports to be (in this case, knowledge of the rules of the road). Thus, if someone leaves the licensing bureau having received a high score on the test and runs a red light not knowing that a red light indicates "stop," we would say that the test is probably not a valid measure of knowledge of the rules of the road. Or, to take another example, if we want to weigh ourselves on scales and with two successive weighings find that there is a 10-pound difference, we would say that the scales are not reliable (although many of us would undoubtedly take the lower weight as the true weight!). In this section, we discuss a number of ways that one can determine rater reliability as well as instrument reliability.

4.7.1. Rater Reliability

The main defining characteristic of rater reliability is that scores by two or more raters or between one rater at Time X and that same rater at Time Y are consistent.

Interrater and Intrarater Reliability. Because these concepts are dealt with in greater detail in chapter 8 on data coding, this section on general reliability provides only a simple introduction. In many instances, test scores are objective and there is little judgment involved. However, it is also common in second language research for researchers to make judgments about data. For example, one might have a dataset from which one wants to extract language

related episodes (LREs), defined as "any part of a dialogue in which students talk about the language they are producing, question their language use, or correct themselves or others" (Swain & Lapkin, 1998, p. 326). We want to make sure that our definition of LREs (or whatever construct we are dealing with) is sufficiently specific to allow any researcher to identify them as such.

Interrater reliability begins with a well-defined construct. It is a measure of whether two or more raters judge the same set of data in the same way. If there is strong reliability, one can then assume with reasonable confidence that raters are judging the same set of data as representing the same phenomenon.

Intrarater reliability is similar, but considers one researcher's evaluations of data, attempting to ensure that the researcher would judge the data the same way at different times—for example, at Time 1 and at Time 2, or even from the beginning of the data set to the end of the data set. To do this, one essentially uses a test–retest method (see sec. 4.7.2); two sets of ratings are produced by one individual at two times or for different parts of the data. Similar to interrater reliability, if the result is high, then we can be confident in our own consistency (see chap. 8 for a discussion of ways to calculate interrater reliability).

4.7.2. Instrument Reliability

Not only do we have to make sure that our raters are judging what they believe they are judging in a consistent manner, we also need to ensure that our instrument is reliable. In this section, we consider three types of reliability testing: test–retest, equivalence of forms of a test (e.g., pretest and posttest), and internal consistency.

Test–Retest. In a test–retest method of determining reliability, the same test is given to the same group of individuals at two points in time. One must carefully determine the appropriate time interval between test administrations. This is particularly important in second language research given the likelihood that performance on a test at one time can differ from performance on that same test 2 months later, because participants are often in the process of learning (i.e., do not have static knowledge). There is also the possibility of practice effects, and the question of whether such effects impact all participants equally. In order to arrive at a score by which reliability can be established, one determines the correlation coefficient[5] between the two test administrations.

[5]A correlation coefficient is a decimal (between 0 and 1) that indicates the strength of relationship between two variables. A high correlation coefficient indicates a strong relationship. Correlations are discussed in chapter 9.

Equivalence of Forms. There are times when it is necessary to determine the equivalence of two tests, as, for example, in a pretest and a posttest. Quite clearly, it would be inappropriate to have one version of a test be easier than the other because the results of gains based on treatment would be artificially high or artificially low, as discussed earlier. In this method of determining reliability, two versions of a test are administered to the same individuals and a correlation coefficient is calculated.

Internal Consistency. It is not always possible or feasible to administer tests twice to the same group of individuals (whether the same test or two different versions). Nonetheless, when that is the case, there are statistical methods to determine reliability; split-half, Kuder-Richardson 20 and 21, and Cronbach's a are common ones. We provide a brief description of each.

Split-half procedure is determined by obtaining a correlation coefficient by comparing the performance on half of a test with performance on the other half. This is most frequently done by correlating even-numbered items with odd-numbered items. A statistical adjustment (Spearman-Brown prophecy formula) is generally made to determine the reliability of the test as a whole. If the correlation coefficient is high, it suggests that there is internal consistency to the test.

Kuder-Richardson 20 and 21 are two approaches that are also used. Although Kuder-Richardson 21 requires equal difficulty of the test items, Kuder-Richardson 20 does not. Both are calculated using information consisting of the number of items, the mean, and the standard deviation (see chap. 9). These are best used with large numbers of items.

Cronbach's a is similar to the Kuder-Richardson 20, but is used when the number of possible answers is more than two. Unlike Kuder-Richardson, Cronbach's a can be applied to ordinal data.

4.8. CONCLUSION

In this chapter, we have dealt with some of the general issues that must be considered in designing a research project, such as the importance of properly identifying, operationalizing, and controlling variables, ensuring the internal and external validity of the study, and determining reliability. In the next chapter we deal in greater detail with design.