**Chapter 5:** Quantitative data collection
**Book name:** Research methods in Applied Linguistics
**Writer:** Zoltan Dornyei
**Professor:** Dr.Zoghi M,
**Saeed Mojarradi** Ph.D. Candidate T.3 Tuesday, November 6, 2018

Quantitative data can be obtained in a number of ways. The most common instrument used for this purpose is the test, which has several types (for example, language tests or psychological tests such as aptitude tests or personality batteries).

**Language testing** is a complex and highly specialized issue with an extensive literature, and therefore it would go beyond the scope of this book to cover it. A second form of producing quantitative data is to measure some phenomenon objectively by controlled means (for example, assessing response time or behavioral frequency), and we can also quantify data that was originally collected in a non-quantitative way.

Finally, a frequent method of collecting quantitative data is through conducting a survey using some sort of a questionnaire. Indeed, besides language tests, questionnaires are the most common data collection instruments in applied linguistics.

The second topic concerns experimental and quasiexperi-mental studies, which represent a special research design of data collection and participant manipulation that has been developed—following the model of experiments in the natural sciences— to produce a method of establishing cause—effect relationships.

### 5.1 Sampling in quantitative research

The most frequent question asked by novice researchers before starting a quantitative investigation is 'How many people do I need to include in my study?' In measurement terms this question can be rephrased as 'How large should my sample be?'

And the second question to follow is usually, 'what sort of people shall I select?' —in other words, who shall my sample consist questions reflect the recognition that in empirical research issues of?'

These concerning the participant sample can fundamentally determine the success of a study.

Furthermore, it is also clear that in quantitative research sampling decisions must be taken early in the overall planning process because they will considerably affect the necessary initial arrangements, the timing and scheduling of the project as well as the various costs involved. Let us start our discussion by defining the three concepts that lie at the heart of quantitative sampling: sample, population, and the notion of representativeness.

#### 5.1.1 Sample, population, and representativeness

The sample is the group of participants whom the researcher actually examines in an empirical investigation and the population is the group of people whom the study is about.

For example, the population in a study might be EFL learners in Taiwanese secondary schools and the actual sample might involve three Taiwanese secondary classes. That is, the target population of a study consists of all the people to whom the survey's findings are to be applied or generalized.

A good sample is very similar to the target population in its most important general characteristics (for example, age, gender, ethnicity, educational background, academic capability, social class, or socioeconomic status) as well as all the more specific features that are known to be related to the variables that the study focuses on .

### 5.1.2 Sampling procedures

Broadly speaking, sampling strategies can be divided into two groups:

**(a)** Scientifically sound 'probability sampling', which involves complex and expensive procedures that are usually well beyond the means of applied linguists, and

**(b)** 'non-probability sampling', which consists of a number of strategies that try to achieve a trade-off, that is, a reasonably representative sample using resources that are within the means of the ordinary researcher.

**Probability sampling**

Probability sampling is a generic term used for a number of scientific procedures, the most important of which are the following:

**• Random sampling**

The key component of probability sampling is 'random sampling'. This involves selecting members of the population to be included in the sample on a completely random basis, a bit like drawing numbers from a hat (for example, by numbering each member and then asking the computer to generate random numbers). The assumption underlying this procedure is that the selection is based entirely on probability and chance, thus minimizing the effects of any extraneous or subjective factors. As a result, a sufficiently large sample should contain subjects with characteristics similar to the population as a whole. Although this is rarely fully achieved, the rule of thumb is that random samples are almost always more representative than non-random samples.

**• Stratified random sampling**

Combining random sampling with some form of rational grouping is a particularly effective method for research with a specific focus. In 'stratified random sampling' the population is divided into groups, or 'strata', and a random sample of a proportionate size is selected from each group. Thus, if we want to apply this strategy, first we need to identify a number of parameters of the wider population that are important from the point of view of the research in a 'sampling frame' —an obvious example would be a division of males and females—and then select participants for each category on a random basis.

**• Systematic sampling**

In anonymous surveys it can be difficult to make a random selection because we may have no means of identifying the partici-pants in advance and thus their names cannot be 'put in the hat' (Cohen et al. z000). A useful technical shortcut is in such cases to apply 'systematic sampling', which involves selecting every nth member of the target group.

### 5.1.2 Sampling procedures

Broadly speaking, sampling strategies can be divided into two groups:
**(a)** Scientifically sound 'probability sampling', which involves complex and expensive procedures that are usually well beyond the means of applied linguists, and
**(b)** 'non-probability sampling', which consists of a number of strategies that try to achieve a trade-off, that is, a reasonably representative sample using resources that are within the means of the ordinary researcher.

**Probability sampling**

Probability sampling is a generic term used for a number of scientific procedures, the most important of which are the following:

**•Random sampling**

The key component of probability sampling is 'random sampling'. This involves selecting members of the population to be included in the sample on a completely random basis, a bit like drawing numbers from a hat (for example, by numbering each member and then asking the computer to generate random numbers). The assumption underlying this procedure is that the selection is based entirely on probability and chance, thus minimizing the effects of any extraneous or subjective factors. As a result, a sufficiently large sample should contain subjects with characteristics similar to the population as a whole. Although this is rarely fully achieved, the rule of thumb is that random samples are almost always more representative than non-random samples.

**•Stratified random sampling**
Combining random sampling with some form of rational grouping is a particularly effective method for research with a
Specific focus. In 'stratified random sampling' the population is divided into groups, or 'strata', and a random sample of a proportionate size is selected from each group. Thus, if we want to apply this strategy, first we need to identify a number of parameters of the wider population that are important from the point of view of the research in a 'sampling frame' —an obvious example would be a division of males and females—and then select participants for each category on a random basis.

**•Systematic sampling**
In anonymous surveys it can be difficult to make a random selection because we may have no means of identifying the participants in advance and thus their names cannot be 'put in the hat' (Cohen et al. z000). A useful technical shortcut is in such cases to apply 'systematic
Sampling', which involves selecting every nth member of the target group.

**•Cluster sampling**
One way of making sampling more practical, especially when the target population is widely dispersed, is to randomly select some larger groupings or units of the populations and then examine all the students in those selected units.

**Non-probability sampling**
Most actual research in applied linguistics employs non probability samples.in quantitative research such purposive , non-representative samples may not be seen as a problem, but in quantitative research , which always aims at representativeness , non-probability samples are regarded as less than perfect compromises that reality forces upon the researcher.

**We can distinguish three main non-probabilistic sampling strategies:**

• **Quota sampling and dimensional sampling** Quota sampling' is similar to proportional stratified random sampling without the 'random' element. That is, we start off with a sampling frame and then determine the main proportions of the subgroups defined by the parameters included in the frame. The actual sample, then, is selected in a way as to reflect these proportions, but within the weighted subgroups no random sampling is used but rather the researcher meets the quotas by selecting participants he/she can have access to.

• **Snowball sampling** This involves a 'chain reaction' whereby the researcher identifies a few people who meet the criteria of the particular study and then asks these participants to identify further appropriate members of the population. This technique is useful when studying groups whose member-ship is not readily identifiable (for example, teenage gang members) or when acce.cc to suitable group members is difficult for some reason.

• **Convenience or opportunity, sampling**
The most common sample type in L 2 research is the 'convenience or 'opportunity sample', where an important criterion of sample selection is the convenience of the researcher:

**How large should the sample be?**

When researchers ask the question, 'How large should the sample be?' what ' they really mean is 'How small a sample can I get away with?' Therefore, the often quoted the larger, the better principle is singularly unhelpful for them. Unfortunately, there are no hard and fast rules in setting the optimal 2 sample size; the final answer to the 'how large/small?' question should be the outcome of the researcher considering several broad guidelines:

**• Rules of thumb**
In the survey research literature a range of between one per cent to ten per cent of the population is usually mentioned as the magic sampling fraction, The following rough estimates of sample sizes for specific types of quantitative methods have also been agreed on by several scholars: correlational research—at least 3o participants; comparative and experimental procedures.

**• Statistical consideration**
A basic requirement in quantitative research is that the sample should have a normal distribution, and Hatch and Lazaraton al distrib (1991) argue that to achieve this the sample needs to include 30 or more le people. However, Hatch and Lazaraton also emphasize that smaller sample sizes

can be compensated for by using cent special statistical procedures (for example, non-parametric tests—see Section 9.9).

**• Sample composition**

A further consideration is whether there are any distinct subgroups within the sample that may be expected to behave differently from the others. If we can identify such subgroups in advance (for example, in most studies of school children, girls have been found to perform differently from boys), we should set the sample size so that the minimum size applies to the smallest subgroup in the sample.

**• Safety margin**

When setting the final sample size, it is advisable to leave a decent 'margin' to provide for unforeseen or unplanned circumstances. For example, some participants are likely to drop out of at least some phases of the project; some questionnaires will always have to be disqualified for one reason or another; and we may also detect unexpected subgroups that need to be treated separately.

**• Reverse approach**

Because statistical significance (see Section 9.4.4) depends on the sample size, our principle concern should be to sample enough learners for the expected results to be able to reach statistical significance. That is, we can take a 'reverse approach': first we approximate the expected magnitude or power of the expected results and then determine the sample size that is necessary to detect this effect if it actually exists in the population. For example, at a $p < .05$ significance level an expected correlation of .4o requires at least 25 participants. (These figures can be looked up in correlational tables available in most texts on statistics;

• Researchers invite volunteers to take part in a study (occasionally even offering money to compensate for the4 time).

• The design allows for a high degree of dropout (or `mortality'), in which case participants self-select themselves out of the sample.

• Participants are free to choose whether they participate in a study or not (for example, in postal questionnaire surveys). Self-selection is inevitable to some extent because few investigations can be made compulsory; however, in some cases— for example, in the examples above — it can reach such a degree that there is a good chance that the resulting sample will not be similar to the target population. For example, volunteers may be different from non-volunteers in their aptitude, motivation or some other basic characteristics, and dropouts may also share some common features that will be underrepresented in the sample with their departure (for example, dropouts may be more unmotivated than their peers and therefore their departure might make the remaining participants' general level of motivation unnaturally high).

## 5.2 Questionnaire surveys

Survey studies aim at describing the characteristics of a population by examining a sample of that group. Although survey data can be collected by means of structured interviews (for example, in market research or opinion polls), the main data collection method in surveys is the use of questionnaires; in this section I focus only on this option. The results of a questionnaire survey are typically quantitative, although the instrument may also contain some open-ended questions that will require a qualitative analysis. The main methodological issues concerning surveys are (a) how to sample the participants, which has been discussed above, and (b) how to design and administer the research tool, the 'self-report questionnaire'.

The popularity of questionnaires is due to the fact that they are relatively easy to construct, extremely versatile and uniquely capable of gathering a large amount of information quickly in a

### 5 .2.1 what are questionnaire • s and what do they measure?

Although the term 'questionnaire' is one that most of us are familiar with, it is not a straightforward task to provide a precise definition of it. To start with, the term is partly a misnomer because many questionnaires do not contain any real questions that end with a question mark. Indeed, questionnaires are also often referred to under different names, such as 'inventories', 'forms', `opinionnaires', 'tests', 'batteries', 'checklists', 'scales', 'surveys', 'schedules', `studies', 'profiles', 'indexes/indicators', or even simply 'sheets'. Second, even in the research community the general rubric of 'questionnaire' has been used in at least two broad senses: (a) interview schedules/guides—described in detail in Section 6.4.z; and (b) self-administered pencil-and-paper questionnaires. In this chapter, I cover this second type of questionnaire, defining it as 'any written instruments that present respondents with a series of questions or statements to which they are to react either by writing out their answers or selecting from among existing answers' (Brown zoo': 6).

**What do questionnaires measure?** ) Questionnaires can yield three types of data about the respondent: Factual questions which are used to find out certain facts about the respondents, such as demographic characterisf (for example, age, gender, and race), residential location, marital and socio-economic status, level of educating history, , occupation, language le • un of time spent in an L2. environment, etc. learn in questions which are used to find out what the respondents are p doing or have done in the past, focusing on actions, life-styles, habits, and personal history. Attitudinal questions which are used to find out what people think, covering attitudes, opinions, beliefs, interests, and values.

### 5.2.2 Multi-item scales

One central issue about questionnaire design concerns how the items to be responded to are actually worded. When it comes to assessing non-factual matters such as the respondents' attitudes, beliefs and other personal or mental variables, the actual wording of the items can assume an unexpected importance. Minor differences in how a question is formulated and framed can often produce radically different levels of agreement or disagreement. For example, Converse and Presser (1986: 41) report on a survey in which simply changing 'forbid' to 'not allow' in the wording of the item 'Do you think the United States should [forbid/not allow] public speeches against democracy?'

Significantly more people were willing to 'not allow' speeches against democracy than were willing to 'forbid' them even though 'allow' and 'forbid' are exact logical opposites. Given that in this example only one word was changed and that the alternative version had an almost identical meaning, this is a good illustration that item wording in general has a substantial impact on the responses. So, how can we deal with the seemingly unpredictable impact of item wording?

Do we have to conclude that questionnaires simply cannot achieve the kind of accuracy that is needed for scientific measurement purposes? We would have to if measurement theoreticians — and particularly American psychologist Rensis Likert in his PhD dissertation in 1931 — had not discovered an ingenious way of getting around the problem: by using 'multi-item scales'. These scales refer to a cluster of several differently worded items that focus on the same target. The item scores for the similar questions are summed,

### 5.2.3 Writing questionnaire items

The typical questionnaire is a highly structured data collection instrument, with most items either asking about very specific pieces of information (for example, one's address or food preference) or giving various response options for the respondent to choose from, for example by ticking a box or circling the most appropriate option. This makes questionnaire data particularly suited for quantitative, statistical analysis. It is, of course, possible to devise a questionnaire that is made up of open-ended items (for example, 'Please describe your dreams for the future ...'), thereby providing data that is qualitative and exploratory in nature, but this practice is usually discouraged by theoreticians.

The problem with questionnaires from a qualitative perspective is that they inherently involve a somewhat superficial and relatively brief engagement with the topic on the part of the respondent. Therefore, no matter how creatively we formulate the items, they are unlikely to yield the kind of rich and sensitive description of events and participant perspectives that qualitative interpretations are grounded in. If we are seeking long and detailed personal accounts, other research methods such as interviews (see Section 6.4) are likely to be more suitable for our purpose. Robson (1993: 243) has summarized this point well: 'The desire to use open-ended questions appears to be almost universal in novice researchers, but is usually rapidly extinguished with experience'. Thus, most professional questionnaires are primarily made up of 'closed-ended' items, which do not require the respondents to produce any free writing; instead, respondents are to choose one of the given alternatives (regardless of whether their preferred answer is among them).

The selected response options can, then, easily be numerically coded and entered into a computer database. Having said that, most questionnaires do contain certain partially open-ended items, and these will be summarized after an overview of the most common closed-ended item formats. (For a more detailed description of the item types with illustrations, see Dornyei 2003.)

#### Likert scales

The most famous type of closed-ended items is undoubtedly the 'Liken scale' (named after its inventor), which consists of a characteristic statement and respondents are asked to indicate the extent to which they 'agree' or 'disagree' with it by marking (for example, circling) one of the responses ranging from `strongly agree' to 'strongly disagree'. For example:

Hungarians are genuinely nice people.

Strongly Agree  Agree  Neither agree nor disagree  Disagree  Strongly disagree

After the item has been administered, each response option is assigned a number for scoring purposes (for example, 'strongly agree' = 5 ... 'strongly disagree' = 1) and the scores for the items addressing the same target are summed up or averaged.

#### Semantic differential scales

Another frequently applied way of eliciting a graduated response is the `semantic differential scale'. This technique is popular because by using it researchers can avoid writing statements (which is not always easy); instead, respondents are asked to indicate their answers by marking a continuum (with a tick or an 'X') between two bipolar adjectives at the extremes.

**• Numerical rating scales**

Numerical rating scales' involve 'giving so many marks out of so many', is assigning one of several numbers (which correspond to a series of ordered categories) to describe a feature of the target. We use this technique in every-day life when we say, for example, that on a scale from one to five something (for example, a film) was worth three or four. The popularity of this scaling technique is due to the fact that the rating continuum can refer to a wide range of adjectives (for example, excellent --> poor; conscientious ---> slapdash) )or adverbs (for example, always —> never); in fact, numerical ratings can easily be turned into semantic differential scales and vice versa.

**• Multiple-choice items**

Most applied linguists are familiar with this format because of its popularity in standardized proficiency testing. In questionnaires they are often used when asking about personal information, such as the level of education of the respondents.

**• Rank order items**

It is a common human mental activity to rank order people, objects, or even abstract concepts according to some criterion, and rank order items in questionnaires capitalize on our familiarity with this process. As the name suggests, these items contain some sort of a list and respondents are asked to order the items by assigning a number to them according to their preferences. A short list of, say, three items can then be quantified by assigning three points to the top ranked option, two to the middle and one to the lowest ranked item.

**Open-ended questions**

`Open-ended questions' include items where the actual question is not fol-lowed by response options for the respondent to choose from but rather by some blank space (for example, dotted lines) for the respondent to fill in. As mentioned earlier, questionnaires are not particularly suited for truly qualitative, exploratory research, but some open-ended questions can still have merit. By permitting greater freedom of expression, open-format items can provide a far greater richness than fully quantitative data.

The open responses can offer graphic examples, illustrative quotes, and can also lead us to identify issues not previously anticipated. Furthermore, sometimes we need open-ended items for the simple reason that we do not know the range of possible answers and therefore cannot provide pre-prepared response categories. In my experience, open-ended questions work particularly well if they are not completely open but contain certain guidance, as illustrated by the following four question types:

• Specific open questions ask about concrete pieces of information, such as facts about the respondent, past activities, or preferences.

• Clarification questions can be attached to answers that have some special importance, and such questions are also appropriate after the 'Other' category in a multiple-choice item (for example, by stating 'Please specify ...').

• Sentence completion where an unfinished sentence beginning is presented for the respondents to complete (for example, 'the thing I liked most about the course is ...'). This can elicit a more meaningful answer than a simple question. I have successfully used this technique on various feedback forms.

• Short-answer questions are different from 'essay questions' (which are not recommended in ordinary questionnaires and therefore will not be dis-cussed) in that they are worded in such a

focused way that the question can be answered succinctly, with a 'short answer' —this is usually more than a phrase and less than a paragraph.

**Rules about item wording**

In questionnaire items we should always choose the simplest way to say something. Items need to be kept clear and direct, without any acronyms, abbreviations, colloquialisms, proverbs, jargon, or technical terms. We should try to speak the 'common language' —the best items are the ones that sound like being taken from an actual interview. Avoid ambiguous or loaded words and sentences It goes without saying that any elements that might make the language of the items unclear or ambiguous need to be avoided. The most notorious 'of such elements are: Nonspecific adjectives or adverbs (for example, 'good', 'easy', 'many,' `sometimes', `often'). Items containing universals such as 'all', 'none', 'never'. Modifying words such as 'only', 'just', 'merely'. Words having more than one meaning. Loaded words (for example, 'democratic', 'modern', 'natural', 'free', etc.), because they may elicit an emotional reaction that may bias the answer. It is also obvious that loaded questions such as 'Isn't it reasonable to suppose that ...?' or 'Don't you believe that ...?' are likely to bias the neutral way. Respondent towards giving a desired answer and should be rephrased in Avoid negative constructions Items that contain a negative construction (i.e. including 'no' or 'not') are deceptive because although they read satisfactorily, responding to them can be problematic. For example, what does a negative response to a negative question mean? In order to avoid any possible difficulties, the best solution is to avoid the use of negatives altogether. In most cases negative items can be restated in a positive way by using verbs or express the opposite meaning.

**Avoid double-barreled questions**

'Double-barreled' questions are those that ask two (or more) questions in one while expecting a single answer. For example, the question 'How are your parents?' asks about one's mother and father, and cannot be answered simply if one of them is well and the other unwell. Indeed, questions dealing with pluralisms (children, students) often yield double-barreled questions, but compound questions also often fall into this category (for example, 'do you always write your homework and do it thoroughly?').

With double-barreled questions even if respondents do provide an answer, there is no way of knowing which part of the question the answer responded to. Avoid items that are likely to be answered the same way by everybody In rating scales we should avoid statements that are likely to be endorsed by almost everyone or almost no one. In most cases these items are not
Informative and they are certainly difficult if not impossible to process statistically (because of insufficient variance). Include both positively and negatively worded items In order to avoid a response set in which the respondents mark only one side of a rating scale, it is worth including in the questionnaire both positively and negatively worded items. We should note, however, that it is all too easy to fall into the trap of trying to create a negatively worded item by using some sort of a negative construction (for example, `don't'), which has been previously warned against. Instead, negatively worded items should focus on negative rather than positive aspects of the target (for example, instead of writing 'I don't enjoy learning English' we can write 'Learning English is a burden for me').

### 5.2.4 The format of the questionnaire

Main parts Questionnaires have a fairly standard component structure that consists of the following elements:

• **Title**

Just like any other piece of writing, a questionnaire should have a title to identify the domain of the investigation, to provide the respondent with initial orientation and to activate relevant background knowledge and content expectations.

• **General** introduction

The 'opening greeting' usually describes the purpose of the study and the organization conducting/sponsoring it. Further important functions of this section involve emphasizing that there are no right or wrong answers; promising confidentiality or anonymity and requesting honest answers; and saying 'thank you'.

• **Specific instructions**

These explain and demonstrate (with examples) how respondents should go about answering the questions.

• **Questionnaire items**

• Additional information

At the end of the questionnaire we may include contact name (for example, the researcher's or an administrator's) with a telephone number or address and some explicit encouragement to get a touch if there are any questions. It is a nice gesture (unfortunately too rarely used) to include a brief note promising to send the respondent a summary of the findings if interested. Sometimes questionnaires can also end invitation to volunteer for a follow-up interview.

**Length**

How long is the optimal length of a questionnaire?

It depends on how important the topic of the questionnaire is for the respondent. If we feel very strongly about something, we are usually willing to spend several hours answering questions. However, most questionnaires in applied linguistics concern topics that have a low salience from the respondents' perspective, and in such cases the optimal length is rather short. Most researchers agree that anything that is more than 4-6 pages long and requires over half an hour to complete may be considered too much of an imposition. As a principle, I have always tried to stay within a four-page limit. It is remarkable how many items can be included within four well-designed pages and I have also found that a questionnaire of 3-4 pages does not tend to exceed the 3o-minute completion limit. Layout It is my experience that the layout of the questionnaire is frequently over-looked as an important aspect of the development of the instrument. This is a mistake: over the past zo years I have increasingly observed that producing an attractive and professional design is half the battle in motivating respondents to produce reliable and valid data. After all, as will be discussed in more detail in Section 5.2..6 (on administering the questionnaire), people usually do not mind expressing their opinions and answering questions as long as they think that the survey is serious and professionally conducted. Three points in particular are worth bearing in mind:

• **Booklet format** Not only does the questionnaire have to be short but it also has to look short. I have found that the format that feels most compact is that of a booklet (for example, an A3 sheet folded into two to make a four-page A4 booklet). This format also makes it easy to read and to turn pages (and what is just as important, it also prevents lost pages ...).

• **Appropriate density** With regard to how much material we put on a page, s to be achieved. On the one hand, we want to make the Quantitative data collection pages full because respondents are much more willing to fill in a two-page rather than a four-page questionnaire even if the two instruments have exactly the same number of items. On the other hand, we must not make the pages look crowded (for example by economizing on the spaces separating different sections of the questionnaire). Effective ways of achieving this trade-off involve reducing the margins, using a space-economical font and utilizing the whole width of the page, for example by printing the response options next to each question rather than below it.
• **Sequence** marking I normally mark each main section of the questionnaire with Roman numbers, each question with consecutive Arab figures, and then letter all the subparts of a question.

**Item sequence**
Once all the items to be included in the questionnaire have been written or collected, we need to decide on their order. Item sequence is a significant factor because the context of a question can have an impact on its interpretation and the response given to it. There are four principles in particular that we need to bear in mind:

**Mixing up the scales**

The items from different scales need to be mixed up as much as possible to create a sense of variety and to prevent respondents from simply repeating previous answers.

**Opening questions:** To set the right tone, the opening questions need to be interesting, relatively simple yet at the same time focused on important and salient aspects. We need to be careful not to force the respondents to take a fundamental decision at such an early stage because that would affect all the subsequent answers. So the initial questions need to be relatively mild or neutral.
**Factual** (or 'personal' or 'classification') questions As Oppenheim (199z) concludes, novice researchers typically start to design a questionnaire by putting a rather forbidding set of questions at the top of a blank sheet of paper, asking for name, address, marital status, number of children, religion, and so on. These personal/classification questions resemble the many bureaucratic forms we have to fill in and are best left at the end of the questionnaire. Also, in many cultures issues like age, level of education, or marital status are considered personal and private matters, and if we ask them near the beginning of the questionnaire they might create some resistance in the respondents, or, in cases where respondents are asked to provide their name, this might remind them of the non-anonymous nature of the survey, which in turn may inhibit some of their answers.

### 5.2.5 Developing and piloting the questionnaire

The general importance of piloting research tools and procedures was already highlighted in Section 3.4.1. Because, as we have seen earlier, in questionnaire so much depends on the actual wording of the items, an integral part of questionnaire construction is 'field testing', that is, piloting at various stages of its development on a sample of people who are similar to the target sample for which the instrument has been designed.

The developing and piloting of a questionnaire is a stepwise process:

Drawing up an item pool The first step is to let our imagination go free and create as many potential items for each scale as we can think of —this collection is referred to as the 'item pool'.

In doing so we can draw inspiration/ideas from two sources: (a) qualitative, exploratory data gathered in interviews (one-to-one or focus group) or student essays focusing, on the content of the questionnaire; and (b) established/published questionnaires in the area (borrowing questionnaire items is an acceptable practice if the sources are properly acknowledged).

Initial piloting of the item pool In order to reduce the large list of questions in the item pool to the intended final number, it is useful to first ask 3-4 trusted and helpful colleagues or friends to go through feedback.

Final piloting (dress rehearsal) Based on the feedback received from the initial pilot group we can normally put together a near –final version of the questionnaire that feels satisfactory and that does not have any obvious glitches.
However, we still do not know how the items will work in actual practice that is whether the respondents will reply to the items in the manner intended by the questionnaire designers.
There is only one way to find out: by administering the questionnaire to a group of about 50 respondents who are in every way similar to the target population the instrument was designed for.

Item analysis. The answers of the pilot group are submitted to statistical analysis to fine-tune and finalize the questionnaire.

The procedure usually involves checking three aspects of the response pattern:
**(a)** Missing responses and possible signs that the instructions were not understood correctly;
**(b)** The range of the responses elicited by each item. We should exclude items that are endorsed by almost everyone or by almost no one because they are difficult if not impossible to process statistically (since statistical procedures require a certain amount of variation in the scores); and
**(c)** the internal consistency of multi-item scales._Multi-item scales are only effective if the items within a scale work together in a homogeneous manner, that is, if they measure the same target area. In psychometric terms this means that each item on a scale should correlate with the other items and with the total scale score, which has been referred to as Likert's criterion of 'internal consistency'.

Statistical packages like SPSS (see Section 9.1) offer a very useful procedure, 'reliability analysis', which provides a straightforward technique to exclude items that do not work (see Section 9.3) and to select the best items up to the predetermined length of the instrument. Post hoc item analysis After the administration of the final questionnaire researchers usually conduct a final item analysis following the same procedures as described above to screen out any items that have not worked properly.

### 5.2.6 Administering the questionnaire

One area in which a questionnaire study can go very wrong concerns the procedures used to administer the questionnaire. Strangely enough, this aspect of survey research has hardly ever been discussed in the literature — questionnaire administration is often considered a mere technical issue relegated to the discretion of the research assistants. This is mistaken; there is ample evidence in the measurement literature that questionnaire administration procedures play a significant role in affecting the quality of the elicited responses.

In social research the most common form of administering questionnaires is by mail, but educational research is different in this respect because administration by hand is typically more prominent than postal surveys. In applied linguistic research, group administration is the most

common method of having questionnaires completed, partly because the typical targets of the surveys are language learners studying within institutional contexts, and it is often possible to arrange to administer the instrument to them while they are assembled together, for example, as part of a lesson. In this format a few questionnaire administrators can collect a very large number of data within a relatively short time

The following strategies this objective:

**Advance notice**

Announcing the questionnaire a few days in advance and sending each participant a printed leaflet that invites their participation, explains the purpose and nature of the questionnaire, and offers some sample items is an effective method of generating a positive climate for the administration and of raising the 'professional' feel of the survey.

**Attitudes** conveyed by teachers, parents, and other authority figures Participants are quick to pick up their superiors' (for example, teachers', bosses', parents') attitude towards the survey and only acquiesce if the message the receive is positive. It is therefore an imperative to win the support of all these authority figures in advance. Respectable sponsorship if we represent an organization that is esteem d highly by the respondents, the positive reputation is likely to be projected onto the survey.

**The behavior of the survey administrator**

The administrators of the questionnaire are, in many ways, identified with the whole survey and therefore everything about them matters: their clothes should be business-like but certainly not more formal than what is typical in the given environment; the way they introduce themselves is important: friendliness is imperative and smiling usually breaks the ice effectively.

**Administrator attitudes** A crucial aspect of the survey administrators' behavior is that it should exhibit keen involvement in the project and show an obvious interest in the outcome. Communicating the purpose and significance of the survey.

An important element in selling the survey to the participants is communicating to them the purpose of the survey and conveying to them the potential significance of the results. The introductory speech of the questionnaire administrator needs to be carefully designed to cover the following points: greeting and thanking; the purpose of the survey and its potential usefulness; why the particular participants have been selected; assurance of confidentiality/anonymity; you'. The duration of completing the questionnaire; 'Any questions?'; final 'thank you.

### 5.2.7 Strengths and weaknesses of questionnaires

The main attraction of questionnaires is their efficiency in terms of researcher time and effort and financial resources: by administering a questionnaire to a group of people, one can collect. a huge amount of information in less than an hour, and the personal investment required will be a fraction of what would have been needed for, say, interviewing the same number of people. Furthermore, if the questionnaire is well constructed, processing the data can also be fast and relatively straightforward, especially by using some computer software.

Questionnaires are also very versatile, which means that they can be used successfully with a variety of people in a variety of situations targeting a variety of topics. Respondents usually do

not mind the process of filling in questionnaires and the survey method can offer them anonymity if needed.

As a result of these strengths, the vast majority of research projects in the behavioral and social sciences involve at one stage or another collecting some sort of questionnaire data. On the other hand, questionnaires also have some serious limitations, and it is all too easy to produce unreliable and invalid data by means of an ill-constructed questionnaire.

## 5.3 Experimental and quast-expemental studies

Having talked about one of the most popular quantitative data collection tools, the questionnaire, let us now turn to a special quantitative data col-lection design, the experimental study, which many people would claim to represent quantitative research at its most scientific because it can establish unambiguous cause—effect relationships. Many research studies in applied linguistics are intended to uncover causal links by answering questions such as 'What's the reason for ...?"What happens if/when ...?' and 'What effect does something have on ...?' However, to establish firm cause—effect relationships is surprisingly difficult because in real life nothing happens in isolation and it is hard to disentangle the interferences of various related factors. For example, if we want to compare the effects of using two different course books in language teaching.

### 5.3.1 Quasi-experimental design

In most educational settings random assignment of students by the researcher is rarely possible and therefore researchers often have to resort to a 'quasi-experimental design'. Quasi-experiments are similar to true experiments in every respect except that they do not use random assignment to create the comparisons from which treatment-caused change is inferred (Cook and Campbell 1979).

Because of the practical constraints, working with 'non-equivalent groups' has become an accepted research methodology in field studies where randomization is impossible or impractical. However, in such cases we cannot rely on the neat and automatic way the true experiment deals with various threats to validity but have to deal with these threats ourselves. In practical terms, in order to be able to make causal claims based on a quasi-experimental study, the effects of the initial group-differences need to be taken into account. In a meta-analysis comparing full experimental and quasi-experimental designs,

Heinsman and Shadish (1996) found that if the two research methods were equally well designed, they yielded comparable results. However, the authors also pointed out that it is no easy task to design a quasi-experimental study as well as an experimental study. The results of the meta-analysis pointed to two specific ways of improving the design of quasi-experimental studies:

**(a)** Avoiding any situations whereby the students self-select themselves (for example, volunteer) to be in the treatment group; and
**(b)** Minimizing pre-test differences between the treatment and the control groups as much as possible.

There are two frequently used methods for achieving this latter goal:

**Matching participants in the treatment and control groups**

The most common matching procedure involves equating the control and treatment groups on a case-by-case basis on one or more variables. If we know that some learner characteristics are likely to have an impact on the target variable we are examining in the study (i.e. the 'dependent variable'), we first need to deter-mine or measure the particular individual difference variables (for example, sex or IQ) and then identify participants in the two comparison groups with very similar parameters (for example, a girl with an IQ of around 102 in both groups).

In a quasi-experimental study we are unlikely to be able achieve perfect matching even if we omit some participants who do not have a close counterpart, but the resultant group compositions will be considerably more compatible than the initial one without matching.

Using analysis of covariance (ANCOVA) ANCOVA offers a statistical method for adjusting the post-test scores for any pre-test differences; in other words, we can statistically screen the unwanted effects out of the outcome measure. (The procedure is described in Section 9.7.3•) There is no doubt that a quasi-experimental design leaves a study more vulnerable to threats to validity than a full experimental design, but as Johnson and Christensen (zoo4) emphasize, just because a threat is possible, it does not mean that it is necessarily plausible.

Furthermore, we can significantly reduce this plausibility by applying the above techniques. As a result, it is generally accepted that properly designed and executed quasi-experimental studies yield scientifically credible results.

**5.3. Analyzing the results of experimental and quasi-experimental studies**

Data obtained from a study with a `pre-test—post-test control-group design' can be analyzed in two ways. The simpler method involves first computing `gain scores' separately in the treatment and the control groups by subtracting the pre-test scores from the post-test scores, and then comparing these gain scores by using t-tests or 'analysis of variance' (ANOVA) (see Sections 9.6 and 9.7) to see whether the gain in the treatment condition was significantly bigger than that in the ordinary (control) condition.

A slightly more complicated method is using ANCOVA in which we compare the post-test scores by controlling for the pre-test scores.

Both methods are acceptable and used in contemporary studies, but there is a growing recognition that ANCOVA offers more precise results (Johnson and Christensen zoo4) for at least two reasons:

**First**, several (but not all) methodologists claim that gain scores are not sufficiently reliable as they are systematically related to any random error of measurement.

**Second**, especially in quasi-experimental studies, ANCOVA helps to reduce the initial group differences (as described above).

### 5.3.3 Experimental studies in educational and applied linguistic research

Furthermore, as the authors report, not only has educational intervention research been decreasing in quantity but there has also been a decline in quality, with the interventions becoming shorter. For example, in the four educational psychology journals in 1995, z6 per cent of the interventions lasted more than one day, whereas the same figure in 2004 was only 16 per cent. One important reason for this decline is the increased use of structural equation modelling (see Section 9.10.4), which makes it possible to make 'quasi-causal' claims about outcomes based on non-experimental, correlational research. Thus, by using this procedure, researchers can sometimes (but not always!) circumvent the laborious and lengthy procedure of running a full experimental study. In applied linguistics there was a steady stream of intervention studies in the 196os as part of the 'methods comparison studies' in classroom research (see Section 8.1), but over the subsequent decades experiments became less popular for at least two reasons:

**(a)** many of the topics applied linguists are interested in are not directly related to 'treatment' or 'intervention', that is, they do not easily lend themselves to manipulation (for example, gender differences, personality traits, ethnolinguistic variation); and

**(b)** Experimental research is rather narrow in scope as only one or a few variables can be altered at a time. On the other hand, typical applied linguistic venues such as language classrooms are complex environments where many factors operate simultaneously, and significant changes can often only be achieved if several variables work in concert or in special combinations.

An experimental design, targeting one or two variables, is inadequate to address such multivariate patterns. While these limitations are valid, let me also emphasize that in many situations experimental studies would be feasible.

### 5.3.4 Strengths and weaknesses of experimental and quasi-experimental designs

In order to control all the variables tightly we may end up with artificial frameworks in laboratory conditions which reduce the external validity (i.e. the generalizability) of the study. In other words, experimental studies can sacrifice external validity for enhanced internal validity, which is one reason why the merits of their use in education have been seriously questioned.

And even properly conducted experimental studies can be affected by the Hawthorne effect, in which the outcome of an otherwise solid experiment is not caused by the treatment variable but by the fact that a treatment has been applied, regardless of its nature—.

In quasi-experimental studies we do not have to worry so much about reduced external validity because these investigations take place in authentic learning environments using genuine class groups, but such designs open up the study to a number of new threats due to the inequality of the initial treatment and control groups.

This problem is referred to as the 'selection bias', and it involves outcome differences that are not due to the treatment but are artefacts of the differences in pre-existing characteristics of the groups being compared.

## 5.4 Collecting quantitative data via the Internet

With the growing use of the Internet it was inevitable that researchers would start collecting data by means of web-based procedures. Indeed, given the od, increasingly available and continuously improving computer hardware and software, it is becoming relatively easy to set up an Internet survey or experiment, and a web-based study offers some tempting benefits (for reviews, including information on free software, see Birnbaum 2.004; Fox et al. etas zoo3):

• **Reduced costs** since most universities or research institutes have the necessaries facilities, setting up an Internet-based project is not more expensive tion5 than initiating traditional research, whereas the running costs are significantly lower

• **Convenience of administration**
The main attraction of web-based research is that it does not require administration of the materials in person; once eve1 the recruitment posting had been made, administration is self-running.

• **Automatic coding** Using a so-called `CGI script' to record responses to disc lased makes the coding and recording of the answers automatic, which, coupled )1led, with self-running administration, is a real bonus.

• **High level of anonymity** the perception is that web-based research is truly anonymous, which enhances the level of honesty. However, because the machine number of a submission is traceable and the security of the data Pen during transit over the Internet cannot be guaranteed by the researcher, it is in principle possible to identify respondents by authorities.

• **International access** The Internet does not have boundaries and therefore ups scholars can reach much larger and more diverse samples worldwide than uh would have been possible before. This is good news for cross-cultural research, even though the language barriers are still there.

• **Access to specialized populations** Web-based recruiting allows access to small, scattered, or specialized populations which would otherwise be difficult to reach. Enamored of these benefits and possibilities, an increasing number of investigators has already begun experimenting with web-based research, and it seems that the initial experiences and the public acceptance of the findings have been sufficiently positive for this method to become established. It is not hard to see that web-based research will play a major role in future research.

### 5.4.1 Technical issues

Internet users have different computers, systems, browsers, and monitors, so the actual stimulus received may differ quite a bit from what the investigator has intended. For this reason, in a survey on self-harm, Fox et al. (2003) decided to employ a single HTML page only. authors explain while it may have been possible to improve the user interface , through the use of other technologies (for example, JavaScript and various other plug-ins), it was felt that the benefits would have been outweighed by the potential adverse effects of the questionnaire not being available in the same format to as wide a population. Currently, participation in more complex web-based surveys and experiments is limited by technical issues such as connection speed to the Internet and quality of installed software, but given the speed of progress in these areas such restrictions are likely to be temporary.

### 5.4.2. Sampling issues

The most acute problem associated with Internet-based research is that it is not possible to apply any systematic, purposive sampling strategy. What normally happens in such research is that once the instrument (typically but not always a survey) has been constructed and tested, the investigators contact various Internet discussion groups, bulletin boards, and lists, and/or initiate some sort of snowball sampling by emailing potential participants and then hope for a sizeable sample. This is obviously far from being systematic, but before we decide that this lack of control over who will eventually participate in the study should disqualify such projects from the category of scientific inquiry, we should recall that non-probability sampling is the most common sampling 'strategy' even in non-web-based research—.

So, the main problem does not necessarily concern the unprincipled selection procedures but rather the fact that the actual sample that completes the web-based survey or experiment may be much more heterogeneous than in traditional research, consisting totally of self-selected participants. As a result, even if we have thousands of responses, it may be very difficult to decide how to generalize the findings. While there are no known solutions to the sampling problem, Birnbaum (2004) mentions two strategies that might offer at least a partial solution.

In the first approach we analyze the research question separately within each substratum of the sample (for example, age, education, gender, and other demographic variables). If the same conclusions are reached in each subgroup, this might lend some external validity to the results.

The second approach also involves the comparison of subsamples, but this time the investigators compare the web-based results with the outcomes of a similar, non-web-based survey or experiment. Again, the convergence of the findings can help to validate-the results.

The End
Tuesday, November 6, 2018