3

Quality criteria, research ethics, and other research issues

Before we discuss the collection and analysis of data, we need to address a few basic research issues that have a bearing on our methodological decisions regardless of our paradigmatic choice. First and foremost come the quality criteria for research, because we can only claim that our investigation is indeed a 'disciplined' inquiry if we can set explicit quality standards to achieve. The second topic to cover is research ethics, which is a curiously neglected issue in much applied linguistic research and often surfaces only when researchers realize that they need to produce some 'ethical clearance' for their study if they want to submit the results to obtain a postgraduate degree or to be published in certain journals. Following the analysis of ethical issues we consider the role and significance of research questions and hypotheses—these will be further discussed in Chapter 14. The chapter concludes with the discussion of three topics—piloting the research, research logs, and data management—that are essential issues for launching a research project but which, in my experience, do not receive sufficient attention in applied linguistic research.

3.1 Quality criteria for research

> *'Validity' is another word for truth.*
> (Silverman 2005: 210)

As we have seen, the basic definition of scientific research is that it is a 'disciplined' inquiry, and therefore one thing research cannot afford is to be haphazard or lacking rigour. Accordingly, there is a general consensus amongst researchers that they must continually strive to assess and document the legitimacy of their findings—after all, scholars have to convince their audiences that they should listen to them and, eventually, believe them.

Unfortunately, general agreement about research quality in scholarly circles stops at the recognition of its importance; when it comes to specifying the concrete 'quality criteria' to be applied, the literature is characterized

by a host of parallel or alternative views and very little consensus. The fragmented nature of the domain is well reflected by the fact that there does not even exist any universally accepted terminology to describe quality criteria, and the terms that are more widely known—'validity' and 'reliability' in particular—are subject to ongoing criticism, with various authors regularly offering disparate sets of alternatives. Of course, given the huge importance of research quality, this situation is not surprising: representatives of different research traditions understandably emphasize quality parameters that will allow the type of inquiry they pursue to come out in a good light. The problem is that the scope of possible quality criteria is rather wide—ranging from statistical and methodological issues through real world significance and practical values to the benefits to the research participants—and some parameters that seem to do one research approach justice do not really work well with other approaches, thereby leading to tension between camps and to further proliferation of quality criteria.

I mentioned in Chapter 1 (Section 1.3) that an overview such as this book needs to be rather conservative in its approach; in this spirit, I will centre my discussion around the two best-known relevant concepts, 'validity' and 'reliability', and will use these with regard to both qualitative and quantitative research. We must note, however, that both terms were originally introduced in quantitative research, and therefore while the significance of the two concepts is an unquestionable fact of life in the QUAN paradigm, many QUAL researchers deny the relevance of 'validity' and 'reliability' as defined in quantitative terms. In order to introduce quality criteria that are more suitable for QUAL inquiries, several alternative terms have been proposed: validity has been referred to as 'trustworthiness', 'authenticity', 'credibility', 'rigour', and 'veracity', but none of these have reached a consensus and the terminology in general is a highly debated topic. There have also been attempts to match QUAL and QUAN terms (for example, external validity = transferability; reliability = dependability) but, surely, the whole rationale for developing a parallel terminology is the conviction that there are no straightforward parallels in the two research paradigms. I will describe in Section 3.1.2 the most influential alternative validity typology put forward by Lincoln and Guba (1985), but I will then argue in favour of Maxwell's (1992) well-known taxonomy that is built around facets of qualitative validity.

3.1.1 Quality criteria in quantitative research

Although the previous discussion will have given the impression that the terminology used to describe quantitative quality standards and criteria is relatively unproblematic, this is not entirely the case. The concept of 'reliability' is fairly straightforward, but when we look at 'validity' we find two parallel systems in the quantitative literature—one centred around 'construct validity' and its components, the other around the 'internal/external validity' dichotomy—and scholars tend to be surprisingly vague about the relation-

ship between these two systems: the usual practice is that a work either covers one or the other. This dualistic approach is due to the fact that within the quantitative paradigm meaningfulness has been conceptualized from two perspectives: research design and measurement. (See Bachman 2006; Lynch 2003.)

Research validity concerns the whole research process, and following Campbell and Stanley (1963) focuses on the distinction of 'internal validity', which addresses the soundness of the research (i.e. whether the outcome is indeed a function of the various variables and treatment factors measured), and 'external validity', which concerns the generalizability of the results beyond the observed sample.

Measurement validity refers to the meaningfulness and appropriateness of the interpretation of the various test scores or other assessment procedure outcomes. As we will see below, validity here is seen as a unitary concept, expressed in terms of 'construct validity', which can be further broken down to various facets such as content or criterion validity. It is this perspective (going back to Lado's work on testing) that produced the classic tenet that a test is valid if it measures what it is supposed to measure, even though the current view is that it is neither the instrument, nor the actual score that is valid but rather the interpretation of the score with regard to a specific population.

Thus, the discussion of quantitative quality standards is best divided into three parts: (a) reliability, (b) measurement validity, and (c) research validity.

Reliability

The term reliability comes from measurement theory and refers to the 'consistencies of data, scores or observations obtained using elicitation instruments, which can include a range of tools from standardized tests administered in educational settings to tasks completed by participants in a research study' (Chalhoub-Deville 2006: 2). In other words, reliability indicates the extent to which our measurement instruments and procedures produce consistent results in a given population in different circumstances. The variation of the circumstances can involve differences in administrative procedures, changes in test takers over time, differences in various forms of the test and differences in raters (Bachman 2004b). If these variations cause inconsistencies, or measurement error, then our results are unreliable.

It is important to remember that, contrary to much of the usage in the methodological literature, it is not the test or the measuring instrument that is reliable or unreliable. Reliability is a property of the scores on a test for a particular population of testtakers (Wilkinson and TFSI, 1999) and Bachman (2004b) reminds us that, accordingly, all the professional international standards require researchers to estimate and report the reliability of 'each total score, subscore, or combination of scores that is to be interpreted' (AERA,

APA, and NCME 1999: 31). In the light of this, it is surprising that the vast majority of quantitative social scientists do not provide reliability estimates for their own data (Onwuegbuzie and Leech 2005). This is partly due to the false understanding that reliability is a characteristic of the instrument, which would imply that if we use an instrument that has been documented to produce reliable scores before, we do not need to worry about establishing reliability in our sample again.

Bachman (2004b) offers a detailed description of two general approaches whereby classic test theory provides estimates of reliability: (1) we can calculate the correlation between two sets of scores, for example between two halves of a test or two parallel forms, or two different raters' ratings. (2) We can calculate a statistic known as Cronbach alpha (see Section 9.3), which is based on the variances of two or more scores and serves as an 'internal consistency coefficient' indicating how the different scores 'hang together' (for example, more than two raters' scores or several parallel questionnaire items in a scale).

Measurement validity

As mentioned earlier, the concept of validity from a measurement perspective has traditionally been summarized by the simple phrase: a test is valid if it measures what it is supposed to measure. However, the scientific conceptualization of measurement validity has gone through some significant changes over the past decades. According to Chapelle's (1999) description of the development of the concept in applied linguistics, validity was seen in the 1960s as a characteristic of a language test. Several types of validity were distinguished: 'criterion validity' was defined by the test's correlation with another, similar instrument; 'content validity' concerned expert judgement about test content; and 'construct validity' showed how the test results conformed to a theory of which the target construct was a part. This traditional conceptualization is still pervasive today.

In 1985, the main international guidelines for educational and psychological measurement sponsored by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education—the AERA/APA/NCME Standards for Educational and Psychological Testing (AERA, APA, and NCME 1999)—replaced the former definition of three validities with a unitary concept, 'construct validity'. This change was a natural consequence of the shift from seeing validity as an attribute of the test to considering it the truthfulness of the interpretation of the test scores. Lynch (2003: 149) summarizes the new conception clearly: 'When examining the validity of assessment, it is important to remember that validity is a property of the conclusions, interpretations or inferences that we draw from the assessment instruments and procedures, not the instruments and procedures themselves'. Following the new approach, content- and criterion-related evidence came to be seen as contributors to the overall validity construct along with validity considerations of the consequences of score

interpretation and use, and even reliability was seen as one type of validity evidence. Validity was portrayed now as the conclusion of a complex validity argument which uses various statistical and theoretical sources as evidence.

Thus, construct validity is now generally accepted as an umbrella term, describing a process for theory validation that subsumes specific test validation operations (Smith 2005). These validation operations are carried out within 'validation studies', and can involve both qualitative and quantitative evidence depending on the specific claims put forward and counterclaims rejected in our validity argument (Bachman 2004b). These arguments are never context-free and McNamara (2006) emphasizes that test score inferences need to be revalidated for every major context of use.

To conclude this brief overview of measurement validity, here is a list of four key points provided by Bachman (2004b) based on Linn and Gronlund's work:

- Validity is a quality of the interpretations and not of the test or the test scores.
- Perfect validity can never be proven—the best we can do is provide evidence that our validity argument is more plausible than other potential competing interpretations.
- Validity is specific to a particular situation and is not automatically transferable to others.
- Validity is a unitary concept that can be supported with many different types of evidence.

Research validity

The second type of validity, 'research validity', is broader than measurement validity as it concerns the overall quality of the whole research project and more specifically (a) the meaningfulness of the interpretations that researchers make on the basis of their observations, and (b) the extent to which these interpretations generalize beyond the research study (Bachman 2004a). These two validity aspects were referred to as *internal validity* and *external validity* by Campbell and Stanley (1963) over four decades ago, and although there have been attempts to fine-tune this typology (for example, by Cook and Campbell 1979), the broadly dichotomous system has stood the test of time:

- A research study or experiment has *internal validity* if the outcome is a function of the variables that are measured, controlled or manipulated in the study. The findings of a study are internally invalid if they have been affected by factors other than those thought to have caused them.
- *External validity* is the extent to which we can generalize our findings to a larger group, to other contexts or to different times. A study is externally invalid if the results apply only to the unique sample or setting in which they were found.

In the quantitative paradigm, research validity is demonstrated by ruling out, or providing evidence against, various 'threats' to validity. These threats concern unintended factors, circumstances, flaws or events that can invalidate the results. Internal validity threats involve the use of inadequate procedures or instruments, any unexpected problems occurring during the study or any uncontrolled factors that can significantly modify the results. The main threat to external validity in experimental studies involves any special interaction between our intervention/treatment and some characteristics of the particular group of participants which causes the experiment to work only in our study (for example, the experiment exploits a special feature of the treatment group that is usually not existent in other groups). In survey studies, external validity threats concern inadequate sampling (to be discussed in Chapter 5).

Main threats to research validity

Let us have a look at the six most salient validity threats. I will not divide them into internal and external categories because a flaw in the research design often affects both aspects of research validity.

- *Participant mortality* or *attrition* In studies where we collect different sets of data from the participants (for example, pre- and post-test or multiple tests and questionnaires), subject dropout is always a serious concern. It reduces the size of the sample that has a complete dataset and, what is more worrying from the perspective of validity, the dropout may not be random but differential (i.e. participants who drop out are different from those who stay), leaving the remaining group with disproportionate characteristics.
- The *Hawthorne effect* The term comes from the name of a research site (in Chicago) where this effect was first documented when researchers investigating an electric company found that work production increased when they were present, regardless of the conditions the workers were subjected to. The reason for such an irrational effect is that participants perform differently when they know they are being studied. Mellow *et al.* (1996: 334) found that this threat is particularly salient in applied linguistic research as it may be 'the single most serious threat to studies of spontaneous language use'.
- *Practice effect* If a study involves repeated testing or repeated tasks (for example, in an experimental or longitudinal study), the participants' performance may improve simply because they are gaining experience in taking the particular test or performing the assessed activity.
- *Maturation* While not a concern in short studies such as a one-off survey, participant maturation—that is, physical or mental change with age—can play a major role in longer-term studies. It is inevitable that subjects change in the course of an experiment or between repeated administration of a questionnaire/test due to the passage of time *per se*—the question is how

much this naturally occurring developmental process affects the target variables in a study.

- *Participant desire to meet expectation (social desirability bias)* The participants of a study are often provided with cues to the anticipated results of the project, and as a result they may begin to exhibit performance that they believe is expected of them. A variation of this threat is when participants try to meet social expectations and over-report desirable attitudes and behaviours while underreporting those that are socially not respected.
- *History* Empirical research does not take place in a vacuum and therefore we might be subject to the effects of unanticipated events while the study is in progress. (See, for example, Section 8.4 on the challenges of classroom research.) Such events are outside the research study, yet they can alter the participants' performance. The best we can do at times like this is to document the impact of the events so that later we may neutralize it by using some kind of statistical control.

3.1.2 Quality criteria in qualitative research

The usual statement we find in the literature about quality in QUAL research is that it is less straightforward to define than quality in QUAN research. (For an insightful analysis from an applied linguistic perspective, see Lazaraton 2003.) While in the previous sections we saw that even quantitative research criteria are not completely unambiguous, it is true that setting explicit quality standards in the qualitative paradigm has been particularly problematic—a claim that even qualitative researchers agree with. Sandelowski and Barroso (2002) summarize the situation well:

> Over the past 20 years, reams of articles and books have been written on the subject of quality in qualitative research. Addressing such concepts as reliability and rigor, value and validity, and criteria and credibility, scholars across the practice and social science disciplines have sought to define what a good, valid, and/or trustworthy qualitative study is, to chart the history of and to categorize efforts to accomplish such a definition, and to describe and codify techniques for both ensuring and recognizing good studies. Yet after all of this effort, we seem to be no closer to establishing a consensus on quality criteria, or even agree on whether it is appropriate to try to establish such a consensus.

One reason for these difficulties lies in the fact that although the terms 'validity' and 'reliability' refer to empirical research in general, in practice they have been associated with quantitative methods and their operationalization has traditionally followed quantitative principles. We saw in Chapter 2 that a qualitative study is inherently subjective, interpretive as well as time- and context-bound; that is, in a qualitative inquiry 'truth' is relative and 'facts' depend upon individual perceptions (Morse and Richards 2002); for this

reason several researchers have argued that qualitative research requires its own procedures for attaining validity that are different from those used in quantitative approaches.

The problem is that, although several typologies of 'qualitative validity' have been put forward, none of them have received unequivocal support and therefore at the moment we do not seem to have a straightforward and powerful alternative means of assuring quality and thus confirming the legitimacy of qualitative research. In fact, some scholars have even argued that qualitative research has several built-in criteria for producing valid results by definition, such as the 'thick description' of the targeted phenomenon or the fact that the results are arrived at through an iterative process of going back and forth between the data and the analysis until a 'goodness of fit' is achieved. However, as Lynch (2003: 157) points out, this assumption of automatic quality control 'will not satisfy many people. Evaluation audiences expect explicit evidence in support of validity'. Indeed, the big problem with the 'qualitative-research-is-valid-by-definition' argument is that although it can be made to sound convincing, not all qualitative accounts are equally useful, credible, or legitimate, and in many cases the difference between a more and a less trustworthy account of a phenomenon is not due to the researcher's different perspective but to some methodological factors which distort the results so that these do not reflect the phenomenon in question accurately (Morse and Richards 2002). In short, not every qualitative study is equally stable and correct and therefore we need standards to be able to sift the wheat from the chaff. Let us first look at the 'chaff', that is, some key QUAL quality concerns, before we look at some responses to the quality issue.

Three basic quality concerns in qualitative research

Quantitative researchers sometimes criticize the qualitative paradigm for not following the principles of the 'scientific method' (for example, the objective and formal testing of hypotheses) or having too small sample sizes, but these concerns miss the point as they, in effect, say that the problem with qualitative research is that it is not quantitative enough. There are, however, certain basic quality concerns in QUAL inquiries that are independent of paradigmatic considerations. I have found three such issues particularly salient:

1 *Insipid data* Focusing on 'individual meaning' does not offer any procedures for deciding whether the particular meaning is *interesting* enough (since we are not to judge a respondent's personal perceptions and interpretations), and if it is not sufficiently interesting then no matter how truthfully we reflect this meaning in the analysis we will obtain only low quality results that are 'quite stereotypical and close to common sense' (Seale *et al.* 2004: 2). In other words, the quality of the analysis is dependent on the quality of the original data and I am not sure whether it is possible to develop explicit guidelines for judging one set of complex idiosyncratic meaning as better than another. As Seale *et al.* conclude, past practice in qualitative research

has not always been convincing in this respect, and although taking theoretical sampling seriously does help (see Section 6.2), no qualitative sampling procedure can completely prevent the documentation of the unexciting. This problem is unique to QUAL inquiries because QUAN research addresses the commonality found in larger samples instead of individual meaning.

2 *Quality of the researcher* Morse and Richards (2002) are right when they warn us that any study is only as good as the researcher, and in a qualitative study this issue is particularly prominent because in a way the researcher is the instrument—see Section 2.1.4. This raises a serious question: how can quality criteria address the researcher's skills that are to a large extent responsible for ensuring the quality and scope of the data and the interpretation of the results? Again, quantitative research does not have to face this issue because a great deal of the researcher's role is guided by standardized procedures.

3 *Anecdotalism and the lack of quality safeguards* The final quality concern has been described by Silverman (2005: 211) as follows:

> qualitative researchers, with their in-depth access to single cases, have to overcome a special temptation. How are they to convince themselves (and their audience) that their 'findings' are genuinely based on critical investigation of all their data and do not depend on a few well-chosen 'examples'? This is sometimes known as the problem of anecdotalism.

Indeed, space limitations usually do not allow qualitative researchers to provide more than a few exemplary instances of the data that has led them to their conclusion, a problem that is aggravated by the fact that scholars rarely provide any justification for selecting the specific sample extracts (i.e. do not give any criteria for 'within-case' sampling—see Section 6.2.4). As a result, Miles and Huberman (1994: 2) have concluded: 'We do not really see how the researcher got from 3,600 pages of field notes to the final conclusions, as sprinkled with vivid illustrations as they may be'. Therefore, readers of qualitative studies are usually not in a position to judge the systematicity of the analysis, let alone to produce any possible alternative interpretations. As a consequence, in the absence of any in-built quality safeguards it is unfortunately too easy to abuse the procedure and produce a convincingly qualitative-like report in which the author has chosen a few quotations from a much larger and more complex database that support his/her preconceived argument.

Reliability

Let us start our discussion of specific QUAL quality criteria by examining the notion of 'reliability'. Kirk and Miller (1986) pointed out two decades ago that the main thrust of methodological development in qualitative research had been towards greater validity and therefore reliability had been some-

what overlooked. This situation has not changed greatly over the past two decades.

Reliability refers to the 'degree of consistency with which instances are assigned to the same category by different observers or by the same observer on different occasions' (Silverman 2005: 224). The concept of consistency is also emphasized in Kirk and Miller's (1986: 69) definition of reliability in field work as the degree to which 'an ethnographer would expect to obtain the finding if he or she tried again in the same way', that is, the degree to which 'the finding is independent of accidental circumstances of the research' (p. 20). Morse and Richards' (2002: 168) definition sums up the consistency issue well but at the same time reveals why qualitative reliability has been played down in the past: 'reliability requires that the same results would be obtained if the study were replicated'. The problem is that replication is not something that is easy to achieve in a research paradigm where any conclusion is in the end jointly shaped by the respondents' personal accounts and the researcher's subjective interpretation of these stories. Having said that, it is possible to conduct reliability checks of various sub-processes within a qualitative inquiry, for example of the coding of interview transcripts by asking a second coder to code separately a sizable part of the transcript (either using the researcher's coding template or generating the codes him/herself) and then reviewing the proportion of agreements and disagreements.

Lincoln and Guba's taxonomy of quality criteria

In a spirited denial of the allegations that qualitative research is 'sloppy' and qualitative researchers respond indiscriminately to the 'louder bangs or brightest lights', Lincoln and Guba (1985) introduced the concept of 'trustworthiness' as qualitative researchers' answer to 'validity'. They proposed four components to make up trustworthiness:

a *Credibility*, or the 'truth value' of a study, which is the qualitative counterpart of 'internal validity'.

b *Transferability*, or the 'applicability' of the results to other contexts, which is the qualitative counterpart of 'external validity'.

c *Dependability*, or the 'consistency' of the findings, which is the qualitative counterpart of 'reliability'.

d *Confirmability*, or the neutrality of the findings, which is the qualitative counterpart of 'objectivity'.

These terms are sometimes referred to as 'parallel criteria' because of their corresponding quantitative counterparts. Although these parallel criteria have been embraced by many qualitative scholars, they have also been criticized on several grounds—see Morrow 2005. My personal concern is that the proliferation of terminology is likely to make the picture more confusing; I believe that it is possible to identify standards for qualitative research by

identifying various relevant facets of the traditional concepts of validity and reliability, a practice followed by Maxwell (1992), described below.

Maxwell's taxonomy of validity in qualitative research

In the introduction of his influential typology of validity in qualitative research, Maxwell (1992) explicitly stated that he did not think that QUAL and QUAN approaches to validity were incompatible. He even suggested that his analysis might also have implications for the concept of validity in quantitative and experimental research. Let us examine the five components of his proposed system. (For a critical review, see Winter 2000.)

1 *Descriptive validity* concerns the factual accuracy of the researcher's account. Maxwell (1992) regards this as the primary aspect of validity because all the other validity categories are dependent on it. It refers to what the researcher him/herself has experienced and also to 'secondary' accounts of things that could in principle have been observed, but that were inferred from other data. One useful strategy for ensuring this validity is 'investigator triangulation', that is, using multiple investigators to collect and interpret the data.

2 *Interpretive validity* Descriptive validity was called a primary validity dimension because it underlies all other validity aspects and not because Maxwell (1992) considered descriptiveness the main concern for qualitative research. Instead, he argued, good qualitative research focuses on what the various tangible events, behaviours or objects 'mean' to the participants. Interpretive validity, then, focuses on the quality of the portrayal of this participant perspective. An obvious strategy to ensure this validity is to obtain participant feedback or member checking, which involve discussing the findings with the participants. (For more details, see below.)

3 *Theoretical validity* corresponds to some extent to the internal validity of the research as it concerns whether the researcher's account includes an appropriate level of theoretical abstraction and how well this theory explains or describes the phenomenon in question.

4 *Generalizability* It is interesting that in labelling this category Maxwell (1992) did not use a phrase containing the word 'validity', even though there would have been an obvious term to use: external validity. This is because he further divided 'generalizability' into 'internal generalizability' and 'external generalizability' and this division would not have worked with the term 'external validity' (i.e. we cannot really have 'internal external validity'). Both aspects of generalizability refer to the extension of the account to persons, times or settings other than those directly studied, but 'internal generalizability' concerns generalizing within the community or institution observed, whereas 'external generalizability' refers to generalizing to other communities or institutions.

Duff (2006) points out that many QUAL researchers view the term generalizability suspiciously because it is reminiscent of QUAN methodology, in which the capacity to generalize from the sample to some wider population is one of the key concerns. This is where Maxwell's (1992) distinction between internal and external generalizability is enlightening: he agrees that generalizability plays a different role in QUAL research than it does in QUAN research and therefore internal generalizability is far more important for most qualitative researchers than is external generalizability. He further explains that generalization in qualitative research usually takes place through the development of a theory derived from the particular persons or situations studied which helps to make sense of other situations. In other words, even if the particulars of a study do not generalize, the main ideas and the process observed might. This is why even a single specially selected case can be illuminating. A useful strategy to examine generalizability, recommended by Duff, is to include in the qualitative account the participants' own judgments about the generalizability of the targeted issue/phenomenon.

5 *Evaluative validity* refers to the assessment of how the researcher evaluates the phenomenon studied (for example, in terms of usefulness, practicability, desirability), that is, how accurately the research account assigns value judgments to the phenomenon. Thus, this validity aspect concerns the implicit or explicit use of an evaluation framework (for example, ethical or moral judgements) in a qualitative account, examining how the evaluative claims fit the observed phenomenon. Evaluative validity is gaining importance nowadays with various 'critical' theories becoming increasingly prominent in the social sciences and also in applied linguistics.

Strategies to ensure validity in qualitative research

Throughout this chapter I have been uncertain about how much space to devote to describing the various systems of validity and reliability because these typologies are admittedly not too practical in themselves. However, I agree with Maxwell's (1992) argument that such typologies offer a useful framework for thinking about the nature of the threats to validity and the possible ways that specific threats might be addressed. In the following, I list the most common strategies used to eliminate or control validity threats and to generate trustworthiness.

Building up an image of researcher integrity

It is my conviction that the most important strategy to ensure the trustworthiness of a project is to create in the audience an image of the researcher as a scholar with principled standards and integrity. At the end of the day, readers will decide whether they have confidence in one's research not by taking stock of the various validity arguments but by forming an opinion about the investigator's overall research integrity. This image of integrity is made up of

several small components but there are certain strategies that are particularly helpful in showing up the researcher's high standards (provided, of course, those exist):

- *Leaving an audit trail* By offering a detailed and reflective account of the steps taken to achieve the results—including the iterative moves in data collection and analysis, the development of the coding frames and the emergence of the main themes—researchers can generate reader confidence in the principled, well-grounded and thorough nature of the research process. (This question is further discussed in Section 6.8 on research journals and in Section 13.1.2 on writing up qualitative reports.) As Holliday (2004: 732) summarizes:

  > As in all research, a major area of accountability must be procedure, and when the choices are more open, the procedure must be more transparent so that the scrutinizers of the research can assess the appropriateness of the researcher 's choices.

- *Contextualization and thick description* Presenting the findings in rich contextualized detail helps the reader to identify with the project and thus come on board.

- *Identifying potential researcher bias* Given the important role of the researcher in every stage of a qualitative study, identifying the researcher's own biases is obviously an important issue in an inquiry, and it also creates an open and honest narrative that will resonate well with the audience (Creswell 2003).

- *Examining outliers, extreme or negative cases and alternative explanations* No research study is perfect and the readers know this. Therefore, explicitly pointing out and discussing aspects of the study that run counter to the final conclusion is usually not seen as a weakness but adds to the credibility of the researcher. Similarly, giving alternative explanations a fair hearing before we dismiss them also helps to build confidence in our results.

Validity/reliability checks

The previous strategies did not actually add to the validity of the data but only helped to convince the audience that the study was valid (provided it was). The following strategies are different in that they involve steps taken by the researcher *during* the project to ensure validity, and it is appropriate to start with the popular technique of building into the study various validity checks.

- *Respondent feedback* (or 'respondent validation' or 'member checking') Because of the emphasis placed in qualitative research on uncovering participant meaning, it is an obvious strategy to involve the participants themselves in commenting on the conclusions of the study. They can, for example, read an early draft of the research report or listen to a presenta-

tion of some tentative results or themes, and can then express their views in what is called a 'validation interview'. If there is agreement between the researcher and the participants, the study's validity is indeed reinforced. However, several scholars have raised the question of how any disagreements should be interpreted. In terms of descriptive validity, such checks are undoubtedly useful and as we have seen in the previous section, respondent validation can also enhance generalizability. But whom shall we listen to when there are concerns about the interpretive validity of the results? Even though the participants are clearly the 'insiders', there is no reason to assume that they can interpret their own experiences or circumstances correctly—in family arguments, for example, insiders often have conflicting views about the same phenomena. Thus, the best way of handling such validity checks is to treat them as further data that can contribute to the overall validity argument after proper interpretation.

- *Peer checking* Qualitative studies often include reliability checks performed by peers. They always involve asking a colleague to perform some aspect of the researcher's role—usually developing or testing some coding scheme, but they can also involve performing other activities such as carrying out an observation task—and then comparing the correspondence between the two sets of outcomes. This is a very useful strategy because even low correspondence can serve as useful feedback for the further course of the study, but unfortunately it is often difficult to find someone who is both competent and ready to engage in this time-consuming activity.

Research design-based strategies

Strategies concerning a study's research design can provide the most convincing evidence about the validity of the research as they are an organic part of the project rather than being 'add-ons'. In a way, these practices are not necessarily 'strategic actions' but simply examples of good research practice.

- *Method and data triangulation* The concept of 'triangulation' involves using multiple methods, sources or perspectives in a research project (to be discussed in more detail when describing mixed methods research in Section 3.1.3 and also in Chapter 7). Triangulation has been traditionally seen as one of the most efficient ways of reducing the chance of systematic bias in a qualitative study because if we come to the same conclusion about a phenomenon using a different data collection/analysis method or a different participant sample, the convergence offers strong validity evidence. However, it leaves the same question open as the one already raised with regard to participant feedback: how shall we interpret any emerging disagreement between the corresponding results?

- *Prolonged engagement* and *persistent observation* Research designs that emphasize the quantity of engagement with the target community/phenomenon carry more face validity: most people would, for example, treat an

account by an ethnographer who has spent 15 years studying a community inherently valid, even though this may not necessarily be so (for example, if the observer has 'gone native').

- *Longitudinal research design* Duff (2006) argues that longitudinal studies have the potential to increase the validity of the inferences that can be drawn from them because they can reveal various developmental pathways and can also document different types of interactions over time. (Longitudinal studies are described in Chapter 4.)