COUNCIL OF EUROPE CONSEIL DE L'EUROPE

**December 2004** **DGIV/EDU/LANG (2004) 13**

# Reference Supplement

**to the**

**Preliminary Pilot version of the Manual for**

***Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment***

**Section D: Qualitative Analysis Methods**

Language Policy Division, Strasbourg

**SECTION D**

# QUALITATIVE ANALYSIS METHODS

**Jayanti Banerjee**
**Lancaster University**

Chapter 6 of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF) (henceforth referred to as 'the Manual'), explains that 'internal validation is a pre-requisite for acceptable linking to the CEF' (Council of Europe, 2003: 100). This chapter focuses on how intrinsic test quality might be established by answering questions such as:

i. Are the items really the level(s) they are supposed to be?
ii. Are the results awarded by different raters comparable?
iii. Are subtests supposedly testing different things providing different information?
iv. Are learners focussing on what is being tested or are they focussing on something quite different?
v. Do the interviewers elicit a good performance effectively?

[extracted from Council of Europe, 2003: 100 – 101]

This section of the Reference Supplement is intended to demonstrate how questions about test quality can be answered using qualitative analysis methods. Its content is as follows:

i. An overview of qualitative methods
ii. Verbal reports
iii. Diary Studies
iv. Discourse/conversation analysis
v. Analysis of test language
vi. Data collection frameworks
vii. Task characteristic frameworks
viii. Questionnaires
ix. Checklists
x. Interviews

Sub-sections ii – x have been grouped according to the nature of the data gathered. They will each follow a standard pattern: description of the qualitative method; examples of research using that method; and advice on how to use the method. Where possible, a key reference will be suggested for each method. A full list of references can be found at the end of the section.

Despite the focus of this section upon issues of test quality, I would like to suggest that many of the methods described here could also be used as part of standard-setting procedures. I will return to this in sub-section 6. However, it is important first to understand what each qualitative method entails and how it has been used already in language testing research.

## 1. Qualitative analysis methods

Qualitative approaches to test validation enable test developers and test users to look more closely at how a test is working by focussing on individuals or small groups. They can be distinguished from quantitative approaches in a number of ways. First, as has already been intimated, qualitative approaches focus on individuals or small groups rather than large test populations. Their aim is to gather detailed information about the specific experiences of these individuals or groups. If a quantitative method, such as a large-

scale survey, has revealed a trend, then a qualitative method can be used to explore that trend at the level of the individual – perhaps in order to explain it.

Second, qualitative approaches have been termed 'interactive and humanistic' (see Cresswell, 2003: 181). The involvement with the research participant is closer. This demands a great deal of sensitivity on the part of the researcher. In many cases, the research participants also contribute to the direction of the research.

Third, qualitative research is interpretive and tends to be cyclical and emergent. For instance, if a researcher wanted to explore test administration procedures (in order to check how test secrecy is maintained) they might design a questionnaire to be completed by everyone involved in the administration of the test (teachers, examiners, office staff). They might then decide to interview a selection of respondents in order to explore the answers to certain questionnaire items. Since, the researcher already had answers to the questionnaire, they might go into the interview with a very clear idea of the issues they wished to explore. However, during the interview, the researcher will need to respond to what the respondent says, interpret meaning and judge whether to (and how) to explore unexpected lines of enquiry.

Despite these distinguishing characteristics, however, it is important to view qualitative and quantitative (such as those described in the other sections of this reference supplement) analysis methods as complementary. Each will give you different information about the test that you are validating and will offer an illuminating perspective. Indeed, studies that use qualitative and quantitative methods in this way are increasingly common.

One recent example is a study by Brown (2003) that explored the effect of the interviewer on a test-taker's speaking proficiency. This research developed on an earlier study by Brown & Hill (1998), which used multifaceted Rasch analysis to derive measures of interviewer difficulty. Brown (2003) identified the easiest and most harsh interviewers from this study and selected a candidate that had been interviewed by both these interviewers. Brown & Hill (1998) had established that raters perceived this candidate to be more proficient when she was interviewed by the easy interviewer than when she was interviewed by the more difficult interviewer. Brown (2003) analysed the transcripts of both interviews using conversation analysis (see 3.1, below) in order to understand better the effect of the interviewer on the test-taker's speaking performance. As a result of this analysis, Brown concluded that the 'easy' interviewer provided more support to the test-taker during the speaking test. For instance, she was explicit about what she expected of the test-taker. She also provided feedback that indicated understanding and interest.

Brown (2003) also wished to explore whether the raters' views of this test-taker were affected by the interviewers' behaviour. Therefore, she gathered retrospective verbal reports (see 2.1, below) from 4 of the raters for each of the interviews. Her analysis of the verbal reports shows that the raters paid attention to whether or not the test-taker had produced extended discourse. They consistently judged that the test-taker produced extended discourse more readily with the 'easy' interviewer than with the 'difficult' one.

This combination of quantitative (multifaceted Rasch analysis) and qualitative (conversation analysis and verbal reports) methodology has established that interviewer style/behaviour can affect the speaking score a test-taker receives. It has also explored the features of interviewer style that are particularly influential on candidate performance. This study is useful in demonstrating the complementarity of qualitative and quantitative methodology. The remainder of this section will discuss various qualitative analysis methods, beginning with those that employ the technique of reflection.

## 2. Reflection

Qualitative analysis methods that employ the technique of reflection ask their informants to write or talk about their thought processes and/or actions when preparing for a test, taking test items, reading a test performance, or using a rating scale. Researchers can choose whether or not to be present during the reflection. If the researcher decided that it was not necessary to be present, then it would be more likely that a diary study (see 2.2, below) would be used. Even if they decided to be present, researchers could also decide how much they wish to probe (through interruption at various points or by a post-reflection interview) the informants' reflections. This sub-section will discuss two ways of gathering reflections on test-preparation, test-taking, and assessment processes: verbal reports and diary studies.

### 2.1 Verbal reports

Verbal reports are also referred to as 'verbal protocols'. They are data collected from test takers and/or examiners in which they talk about their thought processes while they take a test or assess a test performance. Verbal reports have been defined in many different ways but the most helpful is probably Green (1998). She defines verbal reports along three parameters:

i. The type of data collected – informants could be asked to speak only their thoughts aloud (a talk aloud) or to also provide other information that is not already in verbal form such as physical movement (a think aloud)
ii. The time lag between the thought or action and the verbalisation – concurrent verbal report or retrospective
iii. The nature of the intervention (if any) – the researcher might ask for explanations of utterances or prompt for more information (mediated) or may remain silent, allowing the informant to report unprompted (non-mediated).

Verbal reports are very useful sources of data about test takers' and/or examiners' processes when taking or assessing tests. However, they are very demanding for informants to provide because you have to perform the test-taking or assessment task and simultaneously talk about what you are doing and thinking. This presents a tremendous cognitive load. It is important, therefore, to train informants in giving verbal reports. The training should be a two-stage process and should be conducted separately with each informant:

**Stage One**

1. Explain what a verbal report is and what is involved.
2. Demonstrate a verbal report. Show the informant an example of a verbal report either by doing one yourself or by playing a video-recording of someone doing a verbal report.

**Stage Two**

Give the informant an opportunity to practice providing verbal reports. Two tasks should be provided, both similar to the tasks that the informant will have to perform for the real data collection. For instance, if you wish to collect verbal report data about a reading test, select two or three items from an equivalent version of the test to use as practice material.

1. Give your informant the first item to complete as a verbal report. Give your informant detailed guidance about the verbal report that is required. If necessary, interrupt during this task to prompt your informant for more information and to make explicit what you would like them to report on.
2. After they have completed the first verbal report practice task, give the informant further feedback (e.g. explain where you would have liked more detail).
3. Then give them the second task. Allow the informant to perform this verbal report under the conditions that you will use for your study.
4. Give the informant more feedback. It is good to tell your informant what you particularly liked about the verbal report they provided. Also explain where you would have liked more detail.

It is important to note that (despite training) some informants are better at giving detailed verbal reports than others. Alderson (1990) investigated the reading comprehension skills used by test-takers when

completing a 10-item academic English reading comprehension test. He conducted verbal reports with two test-takers. Each session lasted approximately one hour and was recorded for transcription and analysis. Alderson (1990) found that one test-taker had considerable difficulty expressing his thoughts while the other seemed to be much more able. He concluded that it is important to identify good informants. I would recommend that you use the training procedure to identify informants who will be comfortable providing a verbal report and who will give you useful data. As Alderson points out, "in qualitative research of this kind, it is more important to identify good informants than to find representative informants" (1990: 468).

It is also important to consider what language the verbal report should be given in. The choice of language is not necessarily straightforward. You and the test-takers might be first language (L1) speakers of Language A but the test might be in Language B. Should you ask the test-takers to give their verbal reports in your shared L1 (Language A) or in the language of the test (Language B)? In some circumstances, you might find that your test-takers speak Language A but you are an L1 speaker of the language of the test, Language B. In this case, should you request that the test-takers use Language B even though it is not their L1? You might like to consider the following issues:

1. Will informants be able to express their thoughts more fully and accurately if they provide verbal reports in their L1 (regardless of the language of the test)?
2. Will it add to the cognitive load experienced by informants if they are taking the test items in one language and providing verbal reports in another language?
3. What would the informants prefer? The test-takers in the Alderson (1990) study both used the language of the test (English) for their verbal reports. When Alderson observed that one test-taker was having great difficulty expressing his thoughts, he encouraged the informant to use his L1. The informant refused because he wished to improve his English (1990: 467).

Once you have identified skilled informants (who can provide verbal reports) and have decided what language you would like to collect the data in, you will need to decide whether or not you would like to collect your data concurrently or retrospectively. Concurrent data has the advantage that you capture the thoughts as they occur, in so far as it is possible to capture instantaneous information about thoughts. However, it is not always easy to collect concurrent data. This can be because of the nature of the task. For instance, it would be very difficult to ask a test-taker to provide a verbal report while they were taking a speaking test. It would prove very difficult to distinguish between the test performance and the verbal report.

The context in which the data is being collected is also important. For instance, it would be difficult to collect concurrent verbal reports during the live administration of a test. The verbal report process might influence the test-taker's performance and this would be unfair if his/her performance were to contribute to an official score.

However, a retrospective report has the disadvantage that the informants' memory of their thoughts during the test-taking or assessment process might be incomplete or inaccurate. Even if the verbal report were collected immediately after the test or rating, informants might forget details of their behaviour. In such circumstances it might be useful to employ 'stimulated recall methodology' (Gass & Mackey, 2000). This is a variation on more traditional retrospective reports because it provides some support for the informant during the recall. This support can take the form of an audio-tape or video recording of the test-taker (taken while they were performing the task) or it can be a copy of their test performance e.g. the written product of an essay task. Gass & Mackey explain that concrete reminders like this will prompt informants to remember the mental processes that occurred during the original activity (2000: 17).

One possible way of using stimulated recall methodology would be through the following two-stage process:

Stage One

The informants view the recording/read their written performance and report on their thoughts at the time that they were taking the test. They should be allowed to stop and/or rewind the tape if they wish.

Stage Two

In the case of an audio or a video-recording, the researcher can play the recording, stopping the tape at various points to probe for further details about the thoughts of the informant at that point in the test. In the case of a written performance, the researcher might want to draw the informant's attention to specific aspects of the text (perhaps certain lexical choices) and probe for further details about how/why the informant made those choices.

It is important to note that stimulated recall methodology need not necessarily be used in conjunction with verbal reports. Gass & Mackey (2000) describe how it might be used in the form of a questionnaire or in a diary study (see sections 5.1 and 2.2 respectively). The key thing to remember is that stimulated recall methodology can be used to support informants when you ask them to provide you with details of their behaviour during tests, their reactions to tests and/or test performances, and their behaviour during the assessment process.

Verbal reports (whether or not in conjunction with stimulated recall methodology) have been used primarily in the areas of reading and writing (both test-taking and assessment). Cohen (1984) used verbal reports to explore the match between the test-taking processes of examinees taking a reading test and the predictions of the test designers. Cohen reported a number of different studies with different groups of students taking different tests. The number of students in each study varied between 22 and 57 and the tests varied in length and composition. Some tests comprised 10 multiple-choice items (based on a single reading passage) while others combined more than one task type (e.g. multiple-choice, short answer questions, cloze passage). The verbal reports in the different studies revealed interesting information about the students' test-taking strategies as well as their test-taking processes. For instance, Cohen reported that students taking the cloze test tended to ignore the test's instruction to read the entire passage before completing any of the blanks (1984: 74).

Alderson's (1990) study also examined test-taker processes in a reading test but had a slightly different aim. He was responding to arguments that reading skills were separable and could be ranked as higher order or lower order. He gathered verbal reports from 2 students. With one student, Alderson gathered a concurrent verbal report. The student voiced his thought processes while he was taking the test. The second student completed the test first. Alderson conducted a retrospective verbal report with this student. Alderson (1990) found that the students did not necessarily use the micro-skills predicted by experts when responding to particular items. His analysis also revealed that it was possible for different test-takers to get an item correct but to arrive at that correct response by different processes. He further found that it was difficult to identify a body of low-order and high-order skills. As a result of this investigation, he questioned whether test developers could state with any confidence what an item in a test was testing.

In the area of writing assessment, verbal report methodology has primarily been used to investigate assessment processes though it has also, as in the case of Cohen (1994), been used to explore how test-takers perform a particular writing task. Cohen's (1994) study encompassed both phases of testing – the test-taking process and the assessment process because he was interested in how summarising tasks work as a testing format. So, he explored the strategies that test-takers use when they have to write a summary and as well as the strategies that assessors use when rating such tasks. His respondents were 5 students (who completed the summary task) and 2 assessors.

Cohen's study was conducted as follows (1994: 177 – 178):

**Test-taker verbal reports**

1. The test-takers were given a two-part test to complete. They were asked to provide verbal reports of their thoughts and their actions while they were taking the test. They were also asked to comment on the input texts they were reading and to describe any difficulties they had in performing the tasks.
2. A researcher observed the test-takers during the test-taking process. She took notes of what the test-takers did while completing the test (all observable strategies) and also intervened when she felt that the test-taker had not reported on an action or had been silent for some time.
3. When they had completed the test, the test-takers were given a questionnaire. This asked them to comment on whether their English course had helped them to perform the summary tasks, their opinion of this test format, their reactions to the presence (and interventions) of the researcher, as well as whether any difficulties they experienced with the summary tasks were due to reading problems or writing problems.
4. All these stages in the study were conducted in the test-takers' L1 (Portuguese).

**Assessor verbal reports**

1. The assessors were asked to provide verbal reports of their thoughts and their actions during the rating process. They were asked to comment on: the way they determined the topic of the input texts, the stages in their rating process, and also to give their views on how well the test-takers had understood the input texts.
2. A researcher was present during the rating process and noted any observable strategies that the raters used.
3. When they had completed the assessment exercise, the raters were given a questionnaire. This asked them to comment on the summary tasks in relation to previous tests of summarising that they had encountered. It also asked the assessors to point out if they had found any aspect of the test difficult to rate and to comment on the test format, the input texts and the scoring procedures.
4. All these stages in the study were conducted in the assessors' L1 (English)

Cohen's (1994) analysis of the resulting data revealed that assessors varied in the criteria that they applied to the summary tasks as well as in the rating procedures they adopted. Cohen concluded that improvements could be made to the reliability of the marking by establishing clear marking procedures and by developing a scoring key (content) for each task. Cohen also found that the test-takers would benefit from training in this task type. Nevertheless, he concluded (1994: 202) that the summary task type was very useful for 'reactivating what [the students] had learnt in their EAP courses'.

Weigle's (1994) research looked at the effect of rater-training on rating processes. Her respondents were 16 raters working on an English as a second language (ESL) placement test of which half were experienced (having been assessors for this test in previous years) and half were inexperienced/new raters. Weigle's study had three main stages (1994: 203 – 204):

**'PRE'**

1. The raters provided background information during an initial interview.
2. They were then given the placement test marking criteria and asked to rate 13 scripts.
3. Following this rating task, the raters were trained in giving verbal reports.
4. The raters practised the verbal report methodology with four scripts for which the scores were known (taken from a previous administration of the test).
5. Finally, the raters were given 13 more scripts (these differed in topic from the scripts assessed as part of step 2) to rate silently.

**'NORM'**

1. Each rater received a 'norming packet' before this stage of the study. The packet contained 10 representative sample compositions that had previously been rated. Each sample had an official score for each subscale on the marking criteria.
2. The raters were required to mark the compositions before attending the norming session and to compare their marks to the officially assigned marks.
3. During the norming session, the raters discussed their scores in order to understand the rationale behind the official score.
4. Each rater was interviewed immediately after the norming session. They were asked for their reactions to the norming session and to comment on what they had learned. They were also asked to discuss the compositions where their judgements had diverged from the official scores.

**'POST'**

1. After the norming session the raters participated in live rating of the placement test. Two weeks after the end of the live rating the raters attended a second interview. At this interview they were first re-trained in verbal reports.
2. They were then given six scripts to mark while practising the verbal report methodology. Four of these scripts were the same scripts they had marked during the 'PRE' stage (step 4, above).
3. After they had completed the verbal reports, the raters were asked to indicate whether or not they had read each essay before. Where they recognised an essay, they were asked if they could remember the scores they had given previously.

All the data-collection sessions (including the norming sessions) were video-recorded. The transcripts of the verbal reports took special note of pauses, false starts and repetitions. Weigle analysed the verbal reports for the four inexperienced raters whose ratings varied the most between the 'PRE' and 'POST' ratings. She found that the rater-training had had two important effects on the ratings that these raters gave. First, they understood the rating criteria better as a result of training. Second, they became more realistic in their expectations of the student performances at each level of ability.

Finally, Lumley (2002) investigated how assessors negotiate their understanding of the rating scale and the test script to arrive at a judgement of the test performance they are rating. The test in question was a high-stakes test that (at the time of data collection) was used as part of the Australian immigration process. Lumley (2002) focussed on 4 experienced assessors, all of whom were accredited raters for this test. His study followed a five-step process (2002: 253):

1. Re-orientation to the rating process (using four practice scripts)
2. Simple rating (no verbal report) (12 scripts of two tasks each)
3. Practice verbal report rating (one practice script)
4. Data collection phase of rating plus concurrent verbal reports (12 scripts of two tasks each)
5. Post-rating interview

You can see from this structure that Lumley employed a well-developed training framework for his assessors, both to re-orient them to the rating process and to familiarise them with the verbal report methodology. Lumley's analysis of the resulting verbal reports revealed the complex relationship between the rater, the writing performance and the rating scale. He was able to identify criteria that the raters used in their judgements but which were not reflected in the rating scale (in this case a criteria relating to the content of the writing – the quantity of ideas) (2002: 263 – 265). He was also able to illustrate how raters negotiate the effect of a test-taker's writing with the criteria in the rating scale, some of which might not be stated explicitly (2002: 265 – 266).

It is more rare to find examples of verbal report methodology in the areas of speaking and listening. This does not mean, however, that such research is not possible. One example is Buck (1994), who used verbal reports for a listening test. At the time that Buck conducted this study he had been unable to find any

published studies using verbal reports with listening comprehension (1994: 153). He therefore conducted a number of pilot sessions in order to explore how best to use verbal report methodology in this context. He conducted the main study with 6 students, all speakers of Japanese. His procedure was as follows:

1. The test-takers took a 54-item test based on a single listening text. The text was divided into 13 short sections that were played to the test-takers one at a time. The items were all short-answer questions. These were divided between the 13 sections. All the questions were in the students' L1 (Japanese) but the students were free to write their responses in either Japanese or English (the language of the test).
2. Each test-taker attended a post-test interview. During this interview, they took the test items again (using the same procedure as adopted during the first administration). But, before they proceeded to each subsequent section, Buck (1994: 154) asked them a number of questions to check how well they had understood the input text and the questions as well as to explore the test-takers' listening and test-taking strategies.

The interviews were conducted in the students' L1 (Japanese) and each lasted approximately two hours. As a result of his analysis of these interviews, Buck concluded that "top-down processes are crucial in listening comprehension" (1994: 163). He also found that listening comprehension was affected by non-linguistic factors such as interest in the subject matter. Listeners make predictions and inferences while listening based on what they have already understood and their background knowledge. Finally, he identified a number of factors that interact to affect student's performance on individual test items.

It is clear from the preceding discussion that verbal reports can offer insights into test quality in a number of ways. These include:

1. The match between test-designers' predictions and the actual skills and processes test-takers use during the test.
2. The role of test-taking strategies in the successful completion of certain task types.
3. The distribution of micro-skills across a test (in order to establish test coverage).
4. An examination of aspects of a particular task-type in order to establish its usefulness in achieving the aims of the assessment.
5. An exploration of what assessors pay attention to and why in order to better understand the effect of these variables on the score that the test-takers receive.
6. The effect of training on what assessors pay attention to and its consequences for inter and intra-rater reliability
7. The effect of the rating scale and rater expertise on what assessors pay attention to.

Though not discussed here, verbal reports could also be used to explore whether and how students' writing processes differ in test and non-test conditions or in different test conditions (such as between paper-based and computer-based tests).

The studies reported here also illustrate some keys points:

1. There is no optimum sample size in a verbal report study. Some studies have involved as few as two respondents while others have involved 50 or more. You will need to judge how many respondents you need in order to be confident that you have captured a healthy range of possible behaviours. However, it is common to have sample sizes of 10 or less.
2. Verbal reports can be gathered for a variety of task-types but you need to bear in mind the length of the data collection session. With the exception of Buck (1994), the sessions reported have been up to 1 hour long. Beyond this you might find that exhaustion sets in and the quality of the verbal report diminishes. If you find that you need to take more time, you might wish to consider breaking the verbal report process into two parts so that you can give your informants a rest period.

3. It is not possible to predict all the directions that the verbal report will take. However, you can increase your preparedness by piloting your methodology.
4. It is usually helpful to combine verbal reports with another type of data collection methodology such as a questionnaire or observation. This will help you to triangulate the information that you gather (i.e. complement it with a view of the same events from another perspective). As a result you may be able to explain more easily what respondents report and/or you might more easily follow up on gaps in the verbal reports.

Despite the potential of this methodology, researchers will inevitably encounter a number of challenges. The first is choosing the context in which to gather the verbal report data. Cohen (1984: 78) argues that you are more likely to capture actual test-taking processes if you gather verbal report data in circumstances when the test result will be official. However, he notes that this places you in something of a 'Catch-22' situation because the students might not be willing to be completely honest. They might worry that a true report of their test-taking processes could adversely affect their mark. Also, as I pointed out earlier, the verbal report process might interfere with the test-taking process and this could also negatively influence the test-takers performance.

The second challenge is ensuring that the verbal reports are sufficiently detailed for profitable analysis. Cohen (1984: 78) points out that verbal reports cannot necessarily capture the level of detail that you might wish for. He gives the example of a multiple-choice item. He explains that, in order to understand fully how one option was selected, you might want the examinee to explain how they eliminated/rejected the alternatives. Yet, despite this attention to detail, Cohen argues that it might not be possible to capture all the processes that occurred in the selection of the answer. Part of this problem, as Alderson (personal communication) suggests, is due to the fact that some processes are simply not accessible to verbal report, perhaps because they occur so quickly and are so automatic that the informant is not aware of them.

Alderson (1990: 477 – 478) also explains that the interviewer might not be aware during the interview of all the areas in the test-taking process that should be probed. As a result, he/she might fail to adequately probe in certain areas at the time and would only realise the gaps during the analysis. He believes that this is due to the reactive nature of the methodology. It is not possible to predict in advance (and therefore be fully prepared for) what will emerge during the verbal report. He suggests that researchers should plan to go back to their informants as soon as possible with follow-up questions and requests for clarification and/or confirmation of interpretations.

One final challenge is that of making sense of the data collected. Buck (1994: 155) points out that the information is often scattered through a number of hours of recordings and it is difficult to decide how best to summarize and present the data in a meaningful form. His solution was to organise his discussion around his initial hypotheses. Cohen adopted a taxonomy developed by Sarig (1987, cited in Cohen, 1994: 179). Unfortunately, there is no single solution to this problem. The approach adopted by one researcher might not be applicable to data gathered in a different context and for a different purpose. As a result each researcher has to find his/her own 'path' through the data collected. Since this conundrum applies to virtually all the methods described in this section of the reference supplement, I will return to it in section 7.5, where I offer some approaches to analysing rich verbal data.

## 2.2 Diary studies

In general, diary studies offer a way of collecting data relatively unobtrusively but regularly. Diary keeping is a familiar activity, even for people who do not keep diaries of their personal lives. It allows researchers to capture people's thoughts and experiences before they can be forgotten or lose their immediacy and significance. However, diaries can vary widely in format. The most familiar format is unstructured, a blank page on which the informant is asked to write everything relating to the area being researched. For instance, a study of how learners prepare for a test and what they focus on might simply

give informants the instruction to write about their daily test-preparation activities. The simplicity of the instruction can result in very interesting and widely varying responses. However, the drawback of providing such an open-ended task is that informants will self-select the information they believe interesting and important. They might provide less data. Alternatively, you might find that the data is extremely varied with the result that if you use an unstructured format with large numbers of respondents, you could find the resulting data very difficult to analyse. It will not have a pre-determined structure and you will have to establish this structure post-hoc.

Symon (1998) argues that most diary studies give their informants more guidance. Some studies can be very structured. They provide informants with diary forms to complete with a combination of closed and open-ended questions (see 5.1 for more discussion of these terms). Respondents have a very clear idea of what they need to include in their diaries and little or no space for including information that has not been explicitly asked for. Taking, once again, the example of a study of how test-takers prepare for a test, a very structured diary entry might list different test preparation activities as a pro-forma. Respondents might then be asked to complete this pro-forma at regular intervals, each time simply ticking the activities they engaged in during the period covered by the pro-forma.

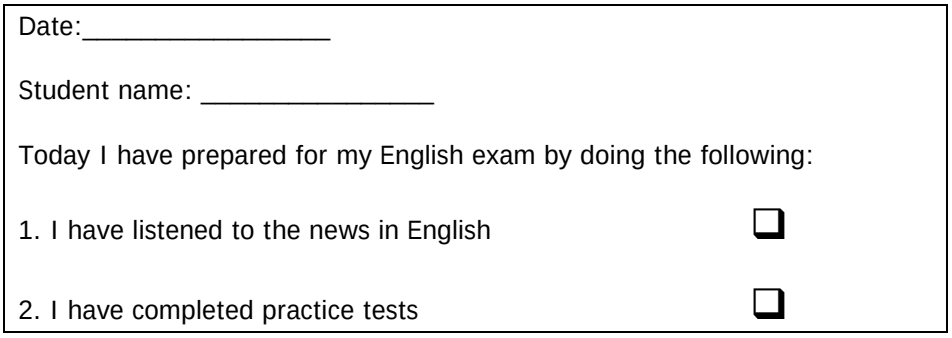


**Figure 1: Excerpt from a structured diary pro-forma**

This approach to diary studies makes analysis very easy because the pro-forma is so structured. It is, therefore, a very good way of using diaries with large numbers of respondents. The problem with providing such strict guidance, however, is that you will only get the information that you ask for. Unless you have been able to successfully predict what your respondents will tell you, a very structured diary form could result in your missing interesting information.

One solution to this is to adopt the middle ground between no guidance and very strict guidance. For instance, if the diary study is of learner strategies when preparing for an examination, it might be possible to give your respondents some examples of test preparation activities that learners might engage in. You could then ask your informants to indicate whether or not they engaged in any of those activities that day or week. You would ask your informants to describe anything else they have done in order to prepare for the examination. You might also ask them to reflect upon how useful they found each of the activities they engaged in. In fact, in order to ensure that your respondents are prompted to provide this additional information (indeed, to check that they are taking the diary study seriously), you should not include the most common test preparation strategies on your initial list. You would expect many of the respondents to add these strategies into their diaries.

As the discussion so far has shown, when deciding on how structured your guidance should be, it is important to think carefully about the purpose of the diary study, the number of respondents you wish to

include in your study and the use that will be made of the data. It is also important to consider a number of other questions:

i. Is the diary study the best way to gather the data? Diary studies provide in-depth, longitudinal data and it is important to decide whether this is appropriate for the research question.
ii. Who is going to complete the diary? Some informants might need more guidance than others – depending on their age and/or their educational level.
iii. What language will the diary be completed in? As in the case of verbal reports, the answer to this question is not always obvious. One consideration might be whether you would like the diary to perform two functions, a research tool for you but also a pedagogic (language learning) tool for your respondents. If you do decide that your respondents should use the diary as a language learning tool, you might wish them to complete it in the target language rather than in their L1.
iv. How often should the diary be completed and for how long? It is particularly important to judge the best time to collect the diaries and this is most successful if the researcher stays in good contact with the informants.
v. How often should you monitor the progress of the diary? Symon (1998: 101) reports that informants are most likely to abandon their diary during the first week of diary-keeping. It is important, therefore, to have frequent contact during that week and then, perhaps, to taper off. However, it is important that contact should be regular.

Though diary studies have not been widely used in published language test validation research, the most common context of use is likely to be learner diaries. Test-takers can be asked to report on their language learning experiences and difficulties post-test. The data collected can be compared to the test score each test-taker was awarded and could provide information about the language abilities of test-takers at different score levels. Other contexts in which diary studies might be useful are examiner/assessor diaries. These could record how markers interpret rating scales and how they apply them to test performances. Diary studies could also be used to explore the behaviour of interlocutors in speaking tests.

## 3. Analysis of samples

Reflections such as verbal reports and diary studies are data that are gathered either after or during test-taking or rating. The next type of qualitative analysis method does not generally involve gathering additional data from test-takers or assessors. Instead, the language of the test becomes the focus of the analysis. In the case of discourse analysis and conversation analysis (see 3.1, below), the test discourse is scrutinised for its social and interactional features. Alternatively, the language of the test can be analysed for features such as grammatical complexity or lexical density (see 3.2, below) perhaps in order to explore whether different tasks tap into different aspects of a test-taker's language resources.

### 3.1 Discourse/conversation analysis

Discourse analysis and Conversation analysis are distinguished from one another in two ways:

1. Discourse analysis is concerned with issues such as power relations and gender inequalities whereas Conversation analysis is more concerned with the extent to which interactions conform to expected patterns.
2. Discourse analysis can be performed on transcripts of conversation or on interviews. It could even be applied to documents (such as test manuals or specifications, perhaps). As Silverman (2001: 178) comments, Discourse analysis is far more 'catholic' about the data it admits. Conversation analysis, however, focuses on transcripts of spoken interaction ('talk').

I will deal with each separately, beginning with Conversation analysis.

Conversation analysis (henceforth CA) is primarily used in the analysis of data from speaking tests. It has three basic assumptions (Heritage, 1984: 241 – 244):

i. Talk has a stable and predictable pattern. The structure of talk can be treated as a 'social fact'.
ii. Each speaker's contribution can only be understood in relation to the context i.e. the preceding sequence of talk. In other words, each utterance inevitably builds on previous utterances and cannot be analysed in isolation from them.
iii. Transcripts must be extremely detailed in order to capture every relevant aspect of speaker meaning because all inference/claims must be grounded in evidence from the data.

CA, therefore, is essentially the analysis of talk in interaction. Hutchby & Wooffitt (1998) provide an excellent introduction to the method. Other good resources are ten Have (1999), Silverman (2001) and Lazaraton (2002). The latter is particularly interesting because it focuses on the use of CA in the validation of speaking tests.

Transcription is a key feature of CA because the transcript must capture as accurately as possible the interaction between the speakers. Hutchby & Wooffitt (1998: 86 – 87) demonstrate the importance of the transcript by presenting two transcriptions of the same conversation. In the first the script simply records what was said, in the order it was said by the two speakers. In the second script, the researcher has indicated where turns overlap and the length of pauses. He/she has also noted other features such as intonation, in-breath, out-breath and emphasis. This transcript shows much more clearly the interaction between the two speakers. It is this transcript that is more helpful in CA. Indeed, because transcripts must be a vivid record of the original interaction, the field has a well-developed glossary of transcription symbols. These can be found in full in Hutchby & Wooffitt (1998: vi – vii). Some of the symbols used are demonstrated in the following example:

```
R: well .hhh let's start with the (0.5) well the MBAs=
I: =yes that sounds fine
R: (1) .hhh Emmanuel=
I: =Emmanuel↑=
R: =yes (.) did the four week course with you:: (0.5)
I: (.) I mean he [was]
R:               [yes] (1) came with first class degree from M ((erased for
confidentiality))=
I: =first class?
R: (1) yes (.) with some experiential learning before that ((reading from
student file)) with business experience before that. (.) this is somebody
who the MBA office asked to do an essay because the experience wasn't so
great (.) they often make sure that the student is understanding .hh is
going to understand what the course is about (.) then they ask them to do
an essay (0.5) and apparently this was a very (3) um (3) this was o.k.::
((laughs))=
I: =right (2) so it wasn't outstanding↑
```

**Figure 2: Example of CA transcription symbols**

| | |
|---|---|
| **(0.5)** | The number in brackets indicates a time gap in tenths of a second |
| **(.)** | A dot enclosed in a bracket indicates a pause in the talk of less than two-tenths of a second |
| = | The 'equals' sign indicates 'latching' between utterances i.e. one utterance follows immediately after the previous one with no break/pause |
| **[ ]** | Square brackets between adjacent lines of speech indicate the onset and end of a spate of overlapping talk |
| **.hh** | A dot before an 'h' indicates speaker in-breath. The more h's the longer the breath |
| **(( ))** | A description enclosed in a double-bracket indicates a non-verbal activity. Alternatively, double brackets may enclose the transcriber's comments |
| **:** | Colons indicate that the speaker has stretched the preceding sound or letter. The more colons the greater the extent of the stretching. |
| **?** | Indicates a rising inflection. It does not necessarily indicate a question. |
| **↓↑** | Pointed arrows indicate a marked falling or rising intonational shift. They are placed immediately before the onset of the shift. |
| Under | Underlined fragments indicate speaker emphasis |

(all taken from Hutchby & Wooffitt, 1998: vi – vii)

The unit of analysis typically is the 'adjacency pair'. An adjacency pair consists of two utterances occurring together that are spoken by two different speakers and function as complementary parts of an exchange. For example:

```
R: well .hhh let's start with the (0.5) well the MBAs=
I: =yes that sounds fine
```

Some common adjacency pairs are:

question – answer
greeting – greeting
invitation – acceptance (refusal)
compliment – acceptance
request – compliance
offer – acceptance (refusal)
complaint – apology

You can see, that the example (above) shows an 'offer-acceptance' adjacency pair. It is important to note, however, that the two parts of an adjacency pair may not be found immediately next to one another. For example:

```
I: =and then what do you do with that book?=
S: =you mean the notebook[?
I:                        [((murmurs agreement))
S:                                  [whenever I have time I just
   take it out and read them.
```

In this example the two sentences in bold are an adjacency pair that is separated by what is called an insertion sequence (another adjacency pair).

CA assumes that these paired (and adjacent utterances) follow certain patterns and rules of interaction. The focus of the analysis is usually on:

1. The structure of the adjacency pair - does the data follow expected patterns such as the ones listed above. How do speakers negotiate breakdowns in the adjacency pairs?)
2. Turn-taking - how speakers negotiate when and for how long they will each speak. This too is believed to be rule governed. In particular, if there is a breakdown in communication or a miscommunication, turn-taking can be inspected and explanations sought.
3. Topic organisation and repair - Test data can be analysed to see who introduces and controls topics and initiates repair as well as the nature of the topic organisation and repair.

I: =which is a fail?=
R: =no (.) actually (.) you're ok .hhh as long as you get over 40% for each module and an average of 50% overall (.) you're ok! he didn't actually fail anything (0.5) I don't remember him doing any re-sits (2) I don't think (.) I didn't keep the breakdown any more than that (2) .hhh so I think he got through every thing in some way (2) but (.) just (.) just overall ((student name)) just seemed to (.) well was quite considerably higher (2) and particularly in the exams as well ((student name)) seemed [to
**I: [57%=**
R: =57% as compared to the 46%↑=
I: =yes (.) but **exams were clearly a problem for both** (1) **so do you think exams place a greater strain on the students' language ability?**=
R: =oh absolutely↑ I think so (.) I think anyone who's doing an exam in a second language (0.5) I mean it's bad enough doing it in your own language but (1) um (1) yes I think (.) you know (.) trying to sort of write under such pressure and such a short time scale and remember everything and be translating it in your head all the time (.) yes. I do.

In this example (above) the excerpts presented in bold type are initiations of topic change. Both topic changes were initiated by 'I', the interviewer. The remainder of the interview could be analysed to establish the extent to which the interviewer initiated the topics that were discussed as well at the extent to which this indicated that the interviewer was in overall control of the interview.

CA has been used by a number of researchers interested in analysing the language of speaking tests. Lazaraton (2002) discusses the use of conversation analysis to analyse test language in the Cambridge EFL examinations. This volume is part of the Studies in Language Testing series published by Cambridge University Press and the University of Cambridge Local Examinations Syndicate. It focuses primarily on CA and includes a number of chapters that explain this analytical approach in detail. In the final chapter, Lazaraton (2002) describes how CA can be used to analyse interviewer behaviour in a speaking test. She presents two studies that were part of the validation programme for the now unavailable Cambridge Assessment of Spoken English (CASE). The data comprised transcripts of test performances for 58 language school students (24 males and 34 females, all Japanese L1 speakers). The performances had been elicited by a pool of 10 examiners. The transcripts were a full record of the elicitations and the student responses. Lazaraton's (2002: 126 – 139) reports the results of these studies, showing how she analysed the transcripts for: the interlocutors' use of the interlocutor frame (which was intended to standardise the input each test-taker received) and also to examine specific aspects of interlocutor behaviour. Her analysis showed that the interlocutors varied widely in their use of the interlocutor frame, using the prompts 40% - 100% of the time. It was also important to note that the same interlocutor would use a different number of prompts in each interview. One interlocutor used between 54% and 77% of the prompts in 6 interviews.

The analysis of specific interlocutor behaviour showed that one interlocutor in particular provided test-takers with supportive behaviour such as:

1. supplying vocabulary
2. rephrasing questions
3. evaluating responses (e.g. 'sounds interesting')
4. repeating and/or correcting responses
5. stating questions that require only confirmation
6. drawing conclusions for candidates

Some interlocutors also used strategies such as 'topic priming' where they first asked a closed question such as 'Do you like to go dancing?' before developing on this with a more open question such as 'What sorts of dancing do you like?'. This too was considered supportive behaviour because it prepared the test-

taker for the upcoming interview question. Supportive behaviour of this kind had a significant effect on the test-takers' performances in one part of the test.

Brown (2003) has also examined the influence of the interviewer upon test-taker performance. She looked in detail at one candidate, who had been interviewed by two different interviewers (in an experimental design). She selected this candidate (Esther) because her scores for the two interviews were markedly different. Indeed, for one interview she was judged as far less able than for the other. Brown (2003) analysed the transcripts of both interviews. She found that one interviewer (Pam) developed on Esther's responses and indicated an interest in what she said, prompting her to elaborate her answers. Pam also used topic primers such as those identified by Lazaraton (2002). Brown (2003) also notes that Pam would close topics consistently i.e. signalling clearly to the test-taker (Esther) that she was about to change to another topic.

Brown's (2003) analysis of the other interviewer (Ian) however, revealed that his behaviour was qualitatively different. Esther had not performed as well when interviewed by Ian as she had when interviewed by Pam. Brown's (2003: 11 – 16) analysis revealed that Ian tended to ask closed questions to which Esther gave short, unelaborated responses. Ian's topic shifts were also more abrupt and did not display the topic priming found in Pam's elicitations. As a consequence, Esther's performance was far less assured. She spoke very little and tended to speak only in short sentences. Brown (2003) argues that the interviewers' behaviour had a clear but unpredictable effect on the test-taker's performance. She concludes that it is very important to examine interviewer behaviour for its possible threat to test validity.

It is clear from the research described above that CA can be used to analyse test language in order to:

1. check the extent to which the test is measuring the desired competences.
2. explore whether test-taker performance is being affected by construct irrelevant factors such as interviewer behaviour.

Like CA, Discourse analysis (henceforth DA) can focus on test performances and need not require the collection of additional data. However, while CA focuses on talk (and therefore is useful in the analysis of the language of speaking tests), DA can also be used to analyse other forms of verbal data such as post-test interviews and test documents e.g. test manuals/handbooks. The other key difference between these two approaches (as mentioned earlier) is the scope of analysis. Whereas CA is primarily concerned with how talk conforms to expected patterns of interaction, DA helps researchers to explore issues such as power relations and gender inequalities. It is defined as the analysis of "texts and talk as social practices" (Potter, 1997: 146) so the analysis focuses on how people use language to 'do' things such as to construct a particular identity or to have a particular effect on their listener. Good introductions to how DA might be performed are provided in Potter & Wetherall (1987), Potter (1996) and Potter (1997) but the use of DA in language testing is best illustrated by examples of research such as Brown & Lumley (1997) Kormos (1999) and O'Loughlin (2002)

Brown & Lumley (1997) studied test-taker performances on the Occupational English Test (OET), a test taken by medical professionals hoping to gain accreditation to practise in Australia. This test consists of two role-plays in which the interlocutor performs the role of a patient or relative of a patient. The test-taker plays their role as the medical professional. The purpose of these role-plays was to simulate, as far as possible, the real situations in which medical professionals need to communicate in order to assess how well the test-takers could cope with these situations. It was important, however, that each test-taker received a comparable level of challenge during the role-plays. Interlocutor variability in the role-plays could undermine the validity of the speaking test.

Consequently, Brown & Lumley's (1997) study explored the behaviour of the interlocutor and its effect on the test-taker's performance (and test score). They analysed test transcripts, paying particular attention

to what the interlocutor said (as part of their role) and the responses they received. The features of interviewer behaviour that appeared to make the test harder were: sarcasm, interruption, repetition (an unwillingness to accept the test-taker's answer to a question), and unco-operativeness. The features of interviewer behaviour that appeared to make the test easier were: the asking of factual questions, linguistic simplification (in the form of repetition of key information, reformulation of key information, slowing of speech etc), and allowing the candidate to initiate topics and to control the interaction.

Brown & Lumley (1997) contended that interlocutors varied in their behaviour depending on the identity they constructed for themselves. An interlocutor who identified with their role as a patient was more likely to produce challenging behaviour whereas an interlocutor who identified more with the test-taker was more likely to produce supportive behaviour. Test-takers who encountered an interlocutor who was more challenging because he/she used sarcasm or was unco-operative had a more difficult test than those who encountered an interlocutor who was generally more supportive. Brown & Lumley (1997) argued that all test-takers should encounter the same level of challenge. In saying this they reminded their readers that this did not preclude the inclusion of some challenging behaviour (for instance, sarcasm) if the construct of the test demanded it. But they contended that if the ability to cope with patient sarcasm should be included as part of the test construct then all the test-takers should receive that challenge.

Kormos (1999) used discourse analysis to examine the effect on the language of the test of different test tasks. She gathered speaking test performances from 30 candidates (10 male and 20 female, all Hungarian L1 speakers). The speaking tests were all conducted by four examiners. Each speaking test comprised three tasks: a general non-scripted interview, a guided role-play, and a picture-description task (1999: 168). Kormos focused particularly on the two interactive tasks – the interview and the role-play. She was interested in exploring the power and dominance relations between the test-taker and the interlocutor in each of these tasks. In order to do this Kormos looked particularly at topic control (topic initiation, ratification and closing) but also looked at how the participants in the speaking test gained the floor (perhaps through interruptions) and retained it. Her analysis revealed a strikingly different pattern of relations between the interview and the role-play. During the interview part of the test, the examiner was dominant. He/she largely had control over the topic (its initiation and closing). The test-taker rejected topics in only 1% of the cases. However, during the role-play task, the test-takers exercised far more control. They initiated 50% more topics than the examiners. During this part of the test, both parties (the test-takers and the examiners) ratified each others' topic initiations 97% of the time. On the basis of this analysis, Kormos (1999) argued that the role-play tasks were a better measure of test-takers' conversational competence because such tasks distributed power more evenly between the candidates and the examiners.

O'Loughlin (2002) was interested in the role of gender on the test-taker's performance and score. His study explored whether there was a gender effect during the interview (in terms of the nature of the interaction between the interlocutor and the candidate) and also during the rating process. He collected test performances from 16 test-takers (8 male and 8 female), each of whom took an International English Language Testing System (IELTS) test twice – once with a female interlocutor and once with a male interlocutor. In the IELTS test, the interlocutor is also the assessor. In addition to the ratings provided by the interlocutor-assessors, O'Loughlin (2002) gathered further ratings of all the test performances from four other assessors (2 male and 2 female). He performed a Rasch analysis of the test scores and a discourse analysis of the test performances. His DA of the test performances focused on three aspects of spoken interaction: overlaps, interruptions and minimal responses. These were chosen because previous research had indicated that these features of spoken interaction were "highly gendered" (O'Loughlin, 2002: 175). O'Loughlin found, however, that there was no clearly gendered pattern in the use of any of the three features he analysed. He conceded that he might have found patterns of gendered language use had he included other features of language in his analysis.

Looking back over these three examples, it is important to note that the research reported here exemplifies the use of DA to analyse speaking tests. This is perhaps the most common use of DA in investigations of test quality. Nevertheless, it is still possible to use DA to analyse other test products such as test manuals or the texts used for reading and listening input.

When using DA to analyse oral language, two important points should be noted. The first is that DA makes use of many of the analytical concepts of CA. For instance, the analysis often focuses on adjacency pairs, turn-taking and topic organisation and repair. Kormos (1999) looked at patterns of topic initiation and uptake while O'Loughlin (2002) looked particularly at how speakers took and held the floor (overlaps, interruptions and minimal responses). The difference, however, is in the perspective taken on the data. In both these cases, the researchers were interested in effect of an aspect of the context or the test-taker upon the patterns of interaction. So Kormos was interested in the effect of the task-type upon the distribution of power in the test discourse and O'Loughlin explored differences in speaker discourse by gender of the test-taker.

The second point to be noted is that, as with CA, DA analyses transcripts of spoken interaction. But, unlike CA transcripts, DA transcripts need not include precise notations of intakes of breath or of each non-verbal contribution (for instance, particles such as 'mm' and 'uh huh'). Instead, they are more likely to use a sub-set of the transcription annotations described above. Particular attention is paid to pauses, para-linguistic behaviour (such as hand movements or the shrugging of shoulders), overlapping speech and emphasis.

It is clear from all the examples provided in this section that Conversation analysis and Discourse analysis have typically been used to analyse spoken test discourse. They can offer insights into speaking test quality in the following ways:

1. The effect of interlocutor behaviour upon the test-taker's performance.
2. An exploration of the influence of test-taker characteristics (such as gender) upon test performance
3. The effect of task-type upon the test-taker's performance.
4. A comparison between test and non-test language in order to establish the extent to which the test has captured relevant aspects of the test-taker's language ability.

The size of the data sample in the research reported here has varied. Brown (2003) focused on just one test-taker and two interlocutors (selecting this from a larger pool of data). Kormos (1999) analysed the performances of 30 test-takers (and four interviewers) each performing two different tasks. O'Loughlin's (2002) dataset comprised 32 performances from 16 test-takers. You will need to judge how much data you will need in order to be confident about the claims you make but it appears that most researchers gather 30 – 60 performances, depending on the depth and focus of their analysis.

Since the language sample is central to CA and DA, the quality of that sample is important. Recording equipment must be in good working condition so that the recording is clear. The transcription stage is also crucial. A lot of useful detail can be lost if the transcription fails to capture it but you can also waste time and resources if you include more information in your transcript than you eventually use. In the case of CA, there is a well-defined transcription system. DA transcriptions can be more flexible (and less detailed) but, because it is not always possible to tell from the outset what aspects of the data will be salient, I would recommend that you perform one practice transcription and analysis in order to identify the precise level of detail that you need to go into in your transcription. Also, be prepared to modify this detail as your analysis proceeds. This means that you will always need to have the original data recordings at hand so that you can refer to them easily should you need to add detail to the transcript or perhaps simply confirm a particular interpretation of the transcript.

Finally, as the example of O'Loughlin (2002) demonstrates, though it is important to be guided by the literature when selecting features to analyse, it is also important to be data-driven i.e. to look for patterns in the data and seek to explain them.

### 3.2 Analysis of test language

The Conversational analysis and Discourse analysis approaches to analysing language samples focus on the social and interactional features of test language. It is also possible to analyse a test-taker's language output (spoken or written) and/or the test input (e.g. a reading text) for a range of linguistic features. This can be useful for a number of reasons. For instance, Kim's (2004: 31) analysis of cross-sectional data from a group of learners indicates that more proficient learners use more subordinate clauses and more phrases in their writing output. This indicates that better-performing students produce grammatically more complex writing and suggests that an analysis of test-taker output might help us to understand better the language features that distinguish one level of performance from another.

Turning to test input, Laufer & Sim (1985) interviewed students in their L1 about their comprehension of L2 academic reading texts. They found that the students needed vocabulary most in order to understand the texts they were reading. Kelly (1991) presents a similar finding in a study of listening comprehension. In this study, advanced language learners in Belgium were asked to transcribe and translate excerpts from British radio broadcasts. The resulting transcriptions and translations were analysed for their errors and Kelly reports that more than 60% of the errors were lexical in nature (i.e. where the meaning of the word had not been understood). These studies indicate that it might be useful to analyse the language of test input in order to better understand sources of test-taker difficulty and to perhaps better estimate the appropriacy of an input text for a particular level of ability – a measure of 'listenability' or 'readability'.

The range of linguistic features that might be investigated include:

1. lexical richness
2. rhetorical structure/functions
3. genre
4. discourse markers
5. grammatical complexity
6. register
7. accuracy

To do this you would first need to identify appropriate measures of the language feature you would like to analyse. This is more complex than it might at first seem. For instance, Read (2001) describes the different considerations involved in measuring lexical richness. It is important to understand how a 'word' is defined. The first key distinction is between 'function' or 'grammatical' words such as *and*, *a*, *to*, and *this* (articles, prepositions, pronouns, conjunctions, auxiliaries etc) and 'content' words such as nouns, verbs, adjectives and adverbs. Taking the age old example:

> The **quick brown fox jumped** over the **lazy dog**

The words highlighted in 'bold' are the content words. The remainder are the function/grammatical words. The other key distinction is that between 'types' and 'tokens'. In vocabulary research, a 'token' is, quite simply, a word used in a text. Therefore, the number of tokens in a text is equal to the number of words in that text. A 'type' however, is a more selective measure. It takes into account only the number of different word forms used in a text. In other words, if a word form is used more than once (e.g. 'the') it will only be counted the first time it is used. More selective again is the term 'lemma'. This is used only in relation to 'content' words and is a super-ordinate term used to describe a base word and all its inflections e.g. *play*, *plays*, *played*, *playing* or *test*, *tests*, *test's*, *tests'*. A 'word family' is a related concept and refers to words that share a common meaning. Read (2001: 19) provides an example:

leak, leaks, leaking, leaked, leaky, leakage, leaker

He explains that even though some of these words have a more metaphorical meaning than others, they are all closely related. Read (2001: 19) does warn, however, that some word families are not as easy to define. For instance the words *socialist* and *socialite* may originate from the same underlying form 'soci-' but they are so distinct in their meaning that they should probably be classed in different word families.

Estimations of lexical richness further involve the calculation of:

1. lexical variation – the variety of different words used, or what might be described as the 'range of expression' (Read, 2001: 200). This is usually measured by calculating the type-token ratio i.e. the number of different words in the text divided by the total number of words in the text. It is important to note here that, because this is a measure of **lexical** variation, researchers focus their *type* measures on 'content' words only rather than also counting 'grammatical'/'function' words such as articles or prepositions.
2. lexical sophistication – the use of low-frequency words such as technical terms or other uncommon words. This is calculated by dividing the number of sophisticated (low frequency) word families in the text by the total number of word families in the text. When calculating this measure, it is usually important to compare the words used to a list of words that the test-takers might be expected to know e.g. by looking at an official vocabulary list for a particular ability level.
3. lexical density – this involves a comparison between the number of grammatical words and the number of content words and is usually calculated by dividing the total number of content (lexical) words by the total number of words in the text.
4. number of lexical errors – this involves counting the number of errors. These errors can take different forms e.g. choosing the wrong word to express a particular meaning, the use of the wrong form of the word, and the stylistically inappropriate use of a word (for instance a very informal word in a formal piece of writing).

All these calculations seem relatively straightforward, but Read (2001: 201) cautions that the results are premised on a number of key decisions. These include, as has already been mentioned (above), decisions about how words might be classified into word families. Other decisions involve deciding whether a word is a content word or a grammatical one and whether multi-word items (such as idioms or phrasal verbs) should be counted as single units. An example is provided from the Slovenian Primary School Leaving Exam (Alderson & Pižorn, 2004: 156) to demonstrate the decisions that need to be made.

Read the text and find out if the statements below the text are true (T), false (F), or not given in the text (NG). Circle the right answer. The example has been done for you.

**AMAZING TIGERS**

Tigers often have to hunt day and night to get enough to eat. You may think that a tiger can easily bring down any animal it goes after. But that's just not true. In fact, most of the time, the tiger's prey gets away. The great cat succeeds just once in 15 to 20 tries. That's why it sometimes doesn't eat for weeks.

A tiger's body is packed with muscles. So it can leap the distance of two cars parked one in front of another. Despite its huge muscled body, a tiger moves very gracefully through the forest. Its claws are mostly hidden in its paws. It glides on its soft, padded feet.

Like other cats, tigers clean their fur with their rough tongues. A tiger's tail is about half the length of its body. Tigers "talk" to other tigers with tails. An upright tail, shaking slowly back and forth, says "Hello". A lowered tail, moving quickly from side to side says "Better be careful".A tail straight back and moving quickly from side to side says "What's happening? I'm excited."

Tigers mark their home ranges with their scent and urine. These markings act as a special kind of communication between tigers. A female scent also lets the males know when she is ready to mate. A tiger roars as a warning to other tigers to stay away.

| | | | | |
|---|---|---|---|---|
| 0 | A tiger can easily catch any animal. | T | (F) | NG |
| 1 | Tigers are good hunters. | T | F | NG |
| 2 | A tiger's tail is the same length as its body. | T | F | NG |
| 3 | Tigers are dangerous to people. | T | F | NG |
| 4 | Tigers communicate with their tails. | T | F | NG |
| 5 | A tiger's tail can show a tiger's excitement. | T | F | NG |
| 6 | Males never know when to approach females. | T | F | NG |
| 7 | Females are not as strong as males. | T | F | NG |

**Figure 3: Example Task No. 4/25 – English**
**(Slovenian Primary School Leaving Exam)**
**Extracted from Alderson & Pizorn (2004: 156)**

Consider the following phrases in the text: *day and night*, *goes after*, *back and forth*, *in fact*, *most of the time*. Would you consider all of these phrases to be multi-word items (which should be counted as single units) or do you think that one or more should be counted as separate words? Similarly, what would you do with the contractions in the text (*that's*, *doesn't*)? Are these single units or are they two separate words? Read (2001: 201) makes clear that there are no 'wrong' answers. It is more important to be meticulous in your recording of the decisions you take and to spend time at the beginning of the analysis setting up the rules that you intend to follow. Read (2001: 201) further suggests the use of corpus analysis tools such as a concordance (perhaps WordSmith). This will list all the words in the text and how frequently they are used. It is also possible to compare the words used in the text with a larger corpus such as the British National Corpus (BNC - http://www.natcorp.ox.ac.uk/). Doing so will reveal the words that might be considered low frequency in relation to the large corpus. To do this you will need to use a corpus analysis tool such as WordSmith (http://www.oup.com/elt/global/isbn/6890/). If you do not have easy access to a corpus of spoken and written language nor to a tool such as WordSmith, you might find it helpful to refer to Leech et al. (2001). This volume presents frequency lists based on an analysis of the BNC. It presents rank-ordered and alphabetical frequency lists for the whole corpus and for various subdivisions (e.g. informative vs. imaginative writing, conversational vs. other varieties of speech). Words are presented according to their grammatical use. For instance, 'round' may be used as a preposition or as an adjective. These two uses of the word 'round' are presented separately.

Even when decisions have been made about how to classify words and phrases it is important to note that other issues might need to be addressed. The first is that lexical variation (the type-token ratio of the lexical words in the text) is affected by the length of the text; it tends to drop as texts get longer. This is particularly problematic when analysing test-taker writing output since some test-takers will inevitably write more than others. Researchers have approached this problem differently. Laufer (1991) decided to take the first 250 words of the scripts that she analysed whereas Arnaud (1984) randomly selected 180 words from test-taker scripts for his analysis.

The second issue that should be addressed is how errors might be treated (quite apart from measuring them as suggested above). For instance, when calculating the lexical variation of a test-taker's writing output, do you wish to take account of all the words that the test-taker has written or only the ones he/she has used correctly? It is also sometimes difficult to decide whether an error is a vocabulary error or a grammatical one. Additionally, it is important to bear in mind that if every error carries the same weight this might skew the results that you get. Therefore, should you ignore minor errors (such as spelling) or should you count every error?

As the foregoing discussion of just one feature has demonstrated, the analysis of test language is a serious undertaking and its exploration requires much preparatory work in order to take defensible decisions. Indeed, the questions that inevitably arise are who is the judge and who has the right to be the judge? Certainly, one way to ensure that your decisions are defensible is to have your categories confirmed by an independent observer (i.e. perform a reliability check) but it is clear that this further lengthens an already complex process. This suggests that it might not be feasible to include analyses of test language as part of your routine checks of test quality. However, as the following descriptions of research will demonstrate, it would certainly be useful if you have a specific question about your test.

O'Loughlin (1995) investigated the comparability of test-taker output in two versions (face-to-face and tape-mediated) of a speaking test. He analysed data gathered from performances on the *Australian Assessment of Communicative English Skills* (henceforth referred to by its acronym - *access:*) comparing the lexical density of the performances on each version. An earlier study by Shohamy (1994) had shown that the language in face-to-face speaking tests (OPIs) tends to contain a higher percentage of grammatical/function words (60% grammatical and 40% lexical words) than the language in tape-mediated speaking tests (SOPIs). This suggests that test-taker output in a SOPI tends to be more 'literate' whereas test-taker output in an OPI tends to be more 'oral'. It further suggests that OPIs and SOPIs do not tap the same underlying construct of speaking. This is of some concern to test developers since they want to ensure that all versions of a test have the same underlying construct. O'Loughlin's (1995) study probed Shohamy's (1994) conclusions by considering the effect of task type on lexical density. The *access:* test was well suited to this exploration because the face-to-face and tape-mediated versions had been developed in parallel and incorporated the same task types.

O'Loughlin's (1995) first step was to develop a comprehensive framework for analysing the test-taker performances (see figure 4, below). Note that O'Loughlin (1995) decided that the verbs 'to be' and 'to have' plus all modals and auxiliaries should count as grammatical items whereas other verbs should be classed as lexical items. Note also his decision to count all contractions as two items (particularly since this was an analysis of speaking output).

O'Loughlin (1995) developed this framework after careful consideration of his data set of 20 speaking performances from 10 test-takers who each took both forms of the *access:* test. He examined this data for the effect on lexical density of both test format (face-to-face or tape-mediated) and task type. To do this, O'Loughlin (1995) focused on four tasks that were roughly parallel in both version of the test – a description, narration, discussion and a role-play. Each task was analysed separately for lexical density. O'Loughlin (1995) was also concerned that his results might differ depending upon the relative frequency of the lexical items used. Therefore, he calculated lexical density using two methods. In the first, he weighted all the lexical items equally regardless of their frequency. In the second, he gave all the high-frequency items half the weighting of the low frequency items.

*A. Grammatical items*
Verbs 'to be' and 'to have'. All **modals** and **auxiliaries**
All **determiners** including articles, demonstrative and possessive adjectives, quantifiers (e.g., some, any) and numerals (cardinal and ordinal).
All **proforms** including pronouns (e.g., she, they, it, someone, something), proverbs (e.g., A: Are you coming with us? B: Yes I *am*), proclauses (e.g., this, that when used to replace whole clauses).
**Interrogative** adverbs (e.g., *what, when, how*) and **negative adverbs** (e.g., *not, never*).
All **contractions**. These were counted as two items (e.g., *they're* = they are) since not all NESB speakers regularly or consistently use contractions.
All **prepositions** and **conjunctions**.
All **dicourse markers** including conjunctions (e.g., *and, but, so*), sequencers (e.g., *next, finally*), particles (e.g., *oh, well*), lexicalised clauses (e.g., *now, then*), spatial deities (e.g., *here, there*) and quantifier phrases (e.g., *anyway, anyhow, whatever*).
All **lexical filled pauses** (e.g., *well, I mean, so*).
All **interjections** (e.g., *gosh, really, oh*).
All **reactive tokens** (e.g., *yes, no, OK, right, mm*).

*B. High-frequency lexical items*
Very common lexical items as per the list of the 700 most frequently used words in English (accounting for 75% of English text) identified in the COBUILD dictionary) project. This list is included in the *Collins COBUILD English course, level 1, student's book*) Willis and Willis, 1988: 111 – 12). It includes **nouns** (e.g., *thing, people*), **adjectives** (e.g., *good, right*), **verbs** (e.g., *do, make, get*), **adverbs of time**, **manner and place** (e.g., *soon, late, very, so maybe, also, too, here, there*). Not items consisting of more than one word are included in this category as the COBUILD list consists of words not items.
**Repetition of low-frequency lexical items** (see below) including alternative word forms of the same item (e.g., *student/study*).

*C. Low-frequency lexical items*
Lexical items not featuring in the list of 700 most frequently used English words cited above including less commonly used **nouns, adjectives, verbs** including participle and infinitive forms (all multiword and phrasal verbs count as one item). Adverbs of **time, place and manner** and all **idioms** (also counted as one item).

**Figure 4: Lexical density – classification of items**
**Taken from O'Loughlin (1995: 228)**

The analyses resulted in data sets comprising percentages of the amount of lexical words/items in the test-takers' output in comparison to the grammatical words/items. Since each test-taker had taken both versions of the test, this meant that there were 8 measures of lexical density for each test-taker. O'Loughlin (1995) reported that the method of calculating the lexical density of test-taker output provided only slightly different results but he argued that the weighted approach was probably more accurate. He also reported that the lexical density of the performances was generally higher for the tape-mediated test. For both test versions, lexical density was lower for the narration task than for the description and discussions tasks. The role-play appeared to be most affected by the test format. In the tape-mediated version, the lexical density was similar to the description and discussion tasks but in the face-to-face version it was lower than all the other tasks analysed. O'Loughlin (1995) concluded that differences between the OPI and the SOPI are more dependent upon the relative interactiveness of the tasks that test-takers are required to perform than upon the test format itself.

Apart from examining the lexical density of two speaking test formats (OPI and SOPI), Shohamy (1994) also conducted a number of other analyses. She first analysed the ideational functions (e.g. describing, elaborating, complaining) of the tasks in the two test formats. She found that the SOPI generally required more functions than all the versions of the OPI analysed i.e. those for low, middle and high level test-takers. Shohamy (1994) then analysed the topics covered by the different versions. She found that low-level test-takers taking the OPI tended to be tested in a narrower range of topics and also on fewer topics. She argued that these results indicated that the OPI implicitly assumed that higher level test-takers were

more able to discuss serious issues. She further argued that because the SOPI presented the same tasks and topics regardless of the level of the test-taker, it gave the test-takers equal opportunities to show what they could do.

Shohamy (1994) then analysed 20 test-taker performances. She calculated the number of errors per performance in relation to the number of words produced, looking particularly at certain error types such as word order, tenses, verb structure and gender. She found that this did not differ significantly between the two test formats. Shohamy (1994) then compared, for each performance, the communicative strategies of shift of topic, hesitation, self-correction, paraphrasing, and switch to L1. She and two independent assessors counted the frequency of occurrence of each of these strategies and then calculated the means for each test performance. The results indicated that paraphrasing was used significantly more frequently in the SOPI. Self-correction also tended to be used more frequently in the SOPI whereas switch to L1 was used more frequently in the OPI.

Shohamy's (1994) final set of analyses compared a number of discourse features of the test-taker performances in each test version. These were:

1. lexical density
2. rhetorical structure of the two test formats
3. genre
4. speech moves e.g. expansion, reporting, description, negotiation for meaning
5. communicative properties e.g. dialogue or monologue, smooth or sharp topic shifts
6. discourse strategies e.g. turn-taking, hesitation, silence
7. content/topics (n.b. this applied the same analyses as had been conducted on the test tasks)
8. prosodic/paralinguistic features e.g. intonation, laughter, hesitations, silence
9. speech functions (n.b. this applied the same analyses as had been conducted on the test tasks)
10. discourse markers e.g. connectors
11. register e.g. level of formality

As a result of this comprehensive analysis, Shohamy (1994) concluded that the SOPI is characterised by concise language that is very similar to a monologue. It is lexically more dense than the OPI and is also more formal. She suggested further that, despite their potential to elicit more functions (as indicated by the analysis of the tasks), the test-taker performances showed that SOPI tasks were more likely to elicit only narrative, reporting and description whereas the OPI had the potential to elicit a wider variety of speech functions. Finally, she argued that the test format could influence the type of language elicited from test-takers.

Wigglesworth (1997) also analysed the language test-takers produced during a tape-mediated speaking test in order to explore the effect of planning time on test-taker output. She was particularly interested in this because the provision of planning time can add considerably to the length of the test. It would also affect the underlying construct of the test. For instance, the question would need to be addressed of whether planning time makes the test more or less authentic. It is therefore important to establish whether such a change to the test is justified by the language that is elicited. Taking a 6-part tape-mediated test, Wigglesworth's (1997) methodology was as follows:

1. She prepared two versions of the test. For both versions two parts (parts 2b and 4) were presented with planning time. For version A planning time was also provided for sections 2a and 3 whereas for version B of the test planning time was provided for sections 2c and 5.
2. She then collected test-performances from 107 test-takers, divided roughly equally between the two test versions.
3. After the test performances had been rated, Wigglesworth (1997) selected a sub-set of 28 performances on each test version dividing these into high and low proficiency candidates.

Once the selected performances had been transcribed, Wigglesworth (1997) divided the texts into clauses. She did this because the dataset was very large and this focus on the clause helped her with the analysis. Wigglesworth (1997) subsequently analysed the texts for:

1. complexity (defined in this case as the number of subordinate clauses used per task)
2. accuracy (i.e. the use of bound morphemes (plural *s*), verbal accuracy, the distribution of definite and indefinite articles)
3. fluency (a type-token analysis was used to measure the number of words used in relation to the number of words used in conjunction with false starts, repetitions and hesitations. The number of clauses containing self-repair was also calculated).

As a result of these analyses, Wigglesworth (1997) reported that high proficiency learners benefited from planning time when performing more difficult tasks. Low proficiency learners did not benefit from planning time on these tasks. She also said that planning time is less beneficial to either group of test-takers when the task is easy, suggesting that this might be because the cognitive load on the students is not heavy in such cases. Her tentative conclusions were that it might be justifiable to provide planning time for complex tasks but not to do so when the tasks were relatively straightforward.

The remaining two examples of research show how analyses of test language can be used to achieve insights into writing test performances. The first, by Ginther & Grant (1997) considered the effects of test-taker ability level and language background and the topic of the task upon the written output. Ginther & Grant analysed 180 exam scripts from the Test of Written English (TWE). Each of these essays had already been rated by two independent assessors using the TWE scale of 1 to 6 where 6 is the highest possible score. The selected scripts had all been given a score of 3, 4 or 5 on the scale (there were insufficient numbers of scripts at the other levels to allow sampling) and represented test-takers with three different L1 backgrounds (Arabic, Chinese and Spanish). Half of the group had written on topic 1 and the other half had written on topic 2.

The essay scripts were then tagged by two independent judges (to allow for a reliability check) for parts-of-speech and for errors. The parts-of-speech coding followed the categories presented in figure 5.

| | |
|---|---|
| Definite article | BE |
| Indefinite article | BE able to |
| Demonstrative adjective | BE going to |
| Adjective | Verb |
| Count noun | Infinitive |
| Noncount noun | Phrasal verb |
| Possessive noun | Preposition |
| Gerund | Multi-word preposition |
| Pronoun | Conjunction |
| Possessive pronoun | Subordinate 1: complement |
| Adverb | Subordinate 2: relative pronoun |
| Multi-word adverb | Subordinate 3: conditional |
| Conjunctive adverb | Subordinate 4: adverbial subordinator |
| Negation | Subordinate 5: present participal subordinator |
| Auxiliary (do/have) | Subordinate 6: wh-interrogative |
| Modal auxiliary | |

**Figure 5: Parts-of-Speech Coding**
**Taken from Ginther & Grant (1997: 388 – 389)**

The categories of error identified were:

1. word form i.e. if the wrong form of a verb, adjective or noun is used (n.b. if there was only one possible correct form, the correct form was also indicated. If there was more than one possible correct answer, then a code was used to indicate this.)
2. word choice e.g. the selection of the wrong preposition
3. word omission .e.g. if the test-taker omitted the article (n.b. omission error codes were placed on the word immediately following the place where the omitted word should have been)
4. spelling

Ginther & Grant (1997) used their analyses to answer the following questions:

1. the influence of test-taker proficiency level on essay errors
2. the influence of test-taker L1 on essay errors
3. the effect of topic on the production of selected parts of speech

They reported that more proficient test-takers (i.e. those who had been rated at level 5 on the TWE scale) wrote longer essays and also produced fewer errors than lower ability test-takers. Additionally, the more proficient test-takers tended to make spelling errors rather than other types of errors whereas the most common error for lower ability test-takers was word form errors. Ginther & Grant also found that the patterns of error by L1 reflected the relative differences or similarities between the test-takers' L1 and English. For instance, the Arab L1 test-takers had the highest percentage of errors per essay and the Spanish L1 speakers the lowest. Chinese and Arabic L1 test-takers were more likely to produce errors of word form whereas the Spanish L1 test-takers most frequently made spelling errors. Interestingly, the Spanish L1 test-takers made more word choice errors than either of the other two L1 groups. Finally, Ginther & Grant (1997) found that the two topics elicited slightly different categories of parts of speech. For instance, topic 1 elicited more examples of negation, gerunds, modal verbs and conditionals than topic 2 whereas topic 2 elicited more adverbs than topic 1. They suggested that this had implications for the equivalence of the topics presented particularly if the mark that the students received was influenced by the presence/absence of certain structures.

Ginther & Grant (1997) suggested a number of avenues for further research. For instance, they said that further analyses should be conducted in order to understand better the effect of certain language features on the marks awarded by assessors. They also suggested that "larger, phrase and sentence-level constructions" should be investigated in order to "evaluate the claim that more complex constructions (such as subordination) are indicative of more mature writers" (1997: 394).

Kim (2004) took a step in this direction in her study of a collection of 33 writing performances by students on an English for Academic Purposes (EAP) course. Her purpose was to describe changes in the grammatical complexity of students' writing that had been placed at different CEF levels. In this small-scale study Kim (2004) focused on three adjacent CEF levels: A2, B1, B2. She conducted three different measures of syntactic complexity:

1. the variety of use of structures
2. the number of subordinate clauses
3. the shift from clauses to phrases

She expected that a comparison of the results of each of these measures would better explain developmental changes between the CEF levels under investigation.

Kim (2004) adopted an analytical framework suggested by Wolfe-Quintero et al (1998), which took the *T-unit* as the basic unit of analysis. The T-unit is also referred to as the terminable unit. It is an independent clause with all its dependent clauses. Take, for example, the following sentence:

> The girl who is getting married tomorrow morning just ran in front of a bus in her haste to collect her wedding dress on time and she was lucky not to be run over.

This sentence comprises two T-units as follows:

i. The girl who is getting married tomorrow morning just ran in front of a bus in her haste to collect her wedding dress on time
ii. She was lucky not to be run over

Kim (2004) conducted the following analyses of each T-unit (ignoring test-takers' errors):

| Measure of syntactic complexity | Analysis |
|---|---|
| variety of use of structures | adverbial clauses per clause (AdcC/C)<br>adjective clauses per clause (AdjC/C)<br>nominal clauses per clause (NoC/C) |
| Number of subordinate clauses | clauses per T-unit (C/T)<br>dependent clauses per T-unit (DC/C)<br>dependent clauses per clause (DC/T) |
| shift from clauses to phrases | prepositional phrases per clause (PP/C)<br>participial phrases per clause (PaP/C)<br>gerund phrases per clause (GP/C)<br>infinitive phrases per clause (IP/C) |

Kim (2004) was then able to compare the analyses for each of the three CEF levels she was investigating. Her results showed a progression from A2 to B2 in all but two of the measures (nominal clauses per clause and gerund phrases per clause). She also found that the results were clearest when comparing A2 and B2. The differences between adjacent levels A2 and B1 were far less clear but there appeared to be a marked increase in syntactic complexity (across measures) when going from B1 to B2.

It is clear from the examples provided in this section that an analysis of test language can provide insights into the:

1. effect of a particular test method upon test-taker performance (for instance, the tape-mediated speaking test)
2. effect of a particular task-type on the language sample elicited
3. influence of topic on the language sample elicited
4. effect of planning time (and other test conditions) upon test-taker performance
5. influence of ability level upon the language sample produced

Unlike CA and DA, an analysis of test language can be performed upon both speaking and writing output. Though no examples have been reported here, I have suggested that it is also possible to analyse the language of the input (for instance in a listening or reading test). I will discuss the analysis of test input in more detail in relation to task characteristic frameworks (see sub-section 4.2, below).

These examples also suggest the following points:

1. The size of the data set can vary. Ginther & Grant (1997) analysed 180 writing scripts while Kim (2004) analysed 33. However, analyses of speaking test language tend to involve relatively small data sets. For instance, O'Loughlin (1995) and Shohamy (1994) studied 20 transcripts of test-taker speaking performances.
2. It is important to define the language features you are using in your analysis. Where competing definitions exist (e.g. O'Loughlin, 1995) I would recommend that you offer a comparison of more than one. In each case, show how the definition affects the results you get and discuss the implications of each for your claims about the quality of your test.

3. Ensure that all your analyses are checked for rater reliability (e.g. Shohamy, 1994; O'Loughlin, 1995; Ginther & Grant, 1997 and Kim, 2004). This will provide proof of the defensibility of your judgements.

Finally, it is important to reiterate that the analysis of language samples is time-consuming and should be used strategically.

## 4. Analytical frameworks

This chapter (particularly section 3) has already made reference to a number of analytical frameworks that can be used to investigate test quality e.g. Conversation Analysis, measures of syntactic complexity and measures of lexical density. Section 5.2 will describe how you might design checklists as a guide for data collection and analysis (usually as part of a study of the test-taking context or of the test-taking process). This section, therefore, will focus on the use of analytical frameworks to analyse test input. The most influential of these is the Framework of Task Characteristics developed by Bachman & Palmer (1996) (see section 4.1, below). However, a recent study involving the CEF has developed a framework that can be used to analyse tests and test specifications (Alderson, personal communication). A brief description of this study is available at http://ling.lancs.ac.uk/groups/ltrg/projects.htm (follow the link for the Dutch CEF construct project).

### 4.1 Task characteristic frameworks

Task characteristic frameworks can help you to analyse your test tasks in some detail in order to explore the extent to which they reflect the test's purpose or perhaps to compare test tasks from two or more versions of a test. The frameworks present a number of 'dimensions' along which the tasks can be analysed or compared. For instance, Weigle (2002: 63) presents a framework that she adapted from Purves et al. (1984: 397 – 8) and Hale et al. (1996) for analysing and comparing writing test tasks. She presents 15 dimensions along which tasks can be described including subject matter, type of stimulus (e.g. graph, table or text), specification of audience, specification of tone, time allowed and choice of prompts.

Fulcher (2003: 57) offers a framework for analysing speaking tasks that includes the following dimensions:

1. Task orientation (for instance is it an open task where the test-taker(s) can decide on the outcome or is the response guided by the rubric? Alternatively, is the task closed and are responses heavily circumscribed?)
2. Interactional relationship (i.e. is there interaction? If there is, how many speakers are involved?)
3. Goal orientation
4. Interlocutor status and familiarity (n.b. in the case of tape-mediated tests it can be argued that there is no interlocutor)
5. Topics
6. Situations

Both Weigle's (2002) and Fulcher's (2003) frameworks are very useful because they are skill specific and therefore take into account characteristics of writing and speaking respectively. A more generic framework is that developed by Bachman & Palmer (1996).

Bachman & Palmer (1996) describe their framework of Task Characteristics as a starting point for task analysis. They list a number of characteristics that should be carefully analysed and described for every task including:

i. the setting (including the physical setting, the participants, and the time of the task)
ii. the test rubrics (including the language of the instructions, the number of parts to the task, the time allotted and the scoring method)
iii. the test input (including the channel of delivery, the length and the characteristics of the language)

iv. the expected response (including the format and the language characteristics)
v. the relationship between the input and the response (including its reciprocity, scope and degree of directness)

(see Bachman & Palmer, 1996: 48 – 57 for more details)

Bachman & Palmer (1996: 57 – 58) suggest that the task characteristics framework can be used as follows:

1. To compare the characteristics of tasks in the target language use situation with test tasks.
2. To analyse existing test tasks in order to make changes or improvements to them.

The Bachman & Palmer framework (as it is commonly referred to) develops on an earlier framework developed by Bachman (1990) called Test Method Facets. This framework was used in a comparison between the Test of English as a Foreign Language (TOEFL) and the Cambridge First Certificate in English (Bachman et al., 1995). Bachman et al. (1995) convened a group of expert judges. These judges were trained to use the framework and subsequently analysed a number of tasks from both tests. The process of training and analysis was as follows:

1. Each judge was given a pair of tests, one from each of the test batteries being studied (FCE and TOEFL). They were asked to study each test carefully and to consider how similar or different they were (and in what ways).
2. The judges were then asked to familiarise themselves with the Test Method Facets framework.
3. They then went through a part of the test, describing it using the Test Method Facets framework. While doing so they were asked to make notes on how well the various descriptive categories in the framework captured their intuitions about the characteristics of the two tests. These notes were used to make revisions to the Test Method Facets framework.
4. The judges then used the revised framework to perform their final analyses of the two tests. For each facet, the judges were asked to place the test task or input text on a three-point scale. For instance, they were asked to rate the rhetorical organisation of the input text on a scale of very simple to very complex. Alternatively they were asked to the number of occurrences of a feature in a test task or input text. For instance, for the facet cultural references, they were asked state whether there were *no occurrences*, *one occurrence* or *two or more occurrences*.

The judges' analyses were used to establish the differences between the tasks on the two tests and to make claims about differences in their underlying constructs. Bachman et al. (1995) reported that agreement between the judges was very high, this implying that the framework helped the experts to pay attention to the key features of the test tasks that were being compared. Their study also demonstrated that the framework allows expert judges to make very detailed judgements about tasks.

However, Clapham (1996) experienced rather more difficulty in applying the Test Method Facets framework in the analysis and comparison of different reading tasks. She tailored the original framework to suit her analysis of IELTS reading tests, reducing the number of facets to 35. However, she found that this was too daunting for her volunteer judges and was forced to reduce the framework further by amalgamating some facets and eliminating others. The final instrument contained only 17 facets. Her procedure consisted of a familiarisation phase and a rating phase. However, despite the familiarisation, Clapham (1996: 149 – 150) remained unsure about their judgements. They commented that some of the categories were not always self-explanatory and were particularly concerned that their analyses would not be stable over time. Finally, Clapham's (1996: 150 – 153) reliability analyses of her judges' ratings revealed quite high agreement for the facets 'grammar' and 'cohesion' but little agreement on facets related to topic specificity. She commented also that her modified Test Method Facets framework did not suit matching and gap-filling tasks (1996: 162).

Any difficulties experienced by researchers are probably because, as Alderson (2000) comments, the framework still needs to be thoroughly investigated through empirical studies and to be modified in the light of the research outcomes. He suggests some possible areas of modification. For instance, the parts of the framework that focus on the characteristics of the test input might not be easy to apply in the analysis of reading test tasks. This is because reading test input comprises both a text and the items that are based upon it. A text might be relatively difficult but the item might be quite straightforward (such as remembering the main 'facts'). Conversely, the text might be quite easy but the item might be rather challenging.

You will have gathered from the discussion so far that there is little published empirical work on the use of task characteristics frameworks. However, despite her own difficulties, Clapham (1996: 162) believes that task characteristics frameworks could be very useful in the content validation of new tests. Indeed, these frameworks have a lot of potential to help us systematise our analyses of test input providing that you bear in mind two guiding principles:

1. You will need to adapt the frameworks already available to suit your test and your context. You will also need to trial and adjust your modified framework until you find that it is practical to use and that your judges understand exactly what they need to do.
2. Remember that the framework is only as good as the judges who use it. Since it is difficult to ensure that a framework is entirely self-explanatory it is important to select your judges carefully and then to familiarise them with the analytical instrument and to also give them sufficient practice in using it before they make 'live' analyses. An issue often debated is whether or not familiarisation and training results in 'cloning' of judgements. This is inevitable and perhaps to some extent some 'cloning' is necessary to ensure the comparability of judgements across raters.

## 5. Feedback methods

Feedback methods such as questionnaires, checklists (particularly observation checklists) and interviews are probably the most familiar methods for gathering qualitative data. They are also typically used in conjunction with each other or with other methods. For instance, in their study of the relationship between students' language proficiency test scores and their subsequent performance on academic degree programmes, Allwright & Banerjee (1997) sent a questionnaire to each student participant at the end of each academic term. The questionnaires were designed to complement each other in order to gather information about each student's study performance and experiences at equally spaced intervals in time. This was to ensure, for example, that the results of the questionnaires at time 2 (in this case the end of the second term of study) could be compared to the results at the end of time 1 (the end of the first term of study) and so on. However, Allwright & Banerjee (1997) also conducted an in-depth interview with each student at the end of their third term of study. During this interview, Allwright & Banerjee (1997) drew on the questionnaire results, probing areas for which the responses had been particularly interesting and also checking their interpretation of the data. They also used the face-to-face meeting to explore aspects of the students' study experiences that were not easy to probe via a questionnaire.

From this example, therefore, it is clear that the different feedback methods are complementary rather than interchangeable. Whenever the circumstances allow, it is often good to 'triangulate' your data by using more than one method (see 7.4 for more discussion). This was the guiding principle behind a set of instruments designed for an International English Language Testing System (IELTS) impact study project (Banerjee, 1996; Herington, 1996; Horák, 1996 and Winetroube, 1997). One set of instruments focused on the classroom. It included a classroom observation schedule, an interview schedule to be used when speaking to the teacher after the observation, and a students' post-observation questionnaire. Further questionnaires were also designed to capture data from teachers and students who were not observed. It is clear from this example that these instruments were intended to complement one another, gathering data from a number of different perspectives and combining different methods of data collection.

The remainder of this section will look more closely at how questionnaires, checklists (including classroom observation schedules) and interviews might be designed.

## 5.1 Questionnaires

Questionnaires gather data that could otherwise also be collected through interviews or focus groups. Their advantage, however, is that they allow researchers to collect views from large numbers of respondents. It can also be easier to manage the data (though this is partly dependent on the questionnaire design) and it is possible to ask face-threatening questions and provide a certain degree of anonymity. Since questionnaires can be completed at any time, respondents also have time to consider their responses.

There are two basic question types – open or closed. Consider the following question pair:

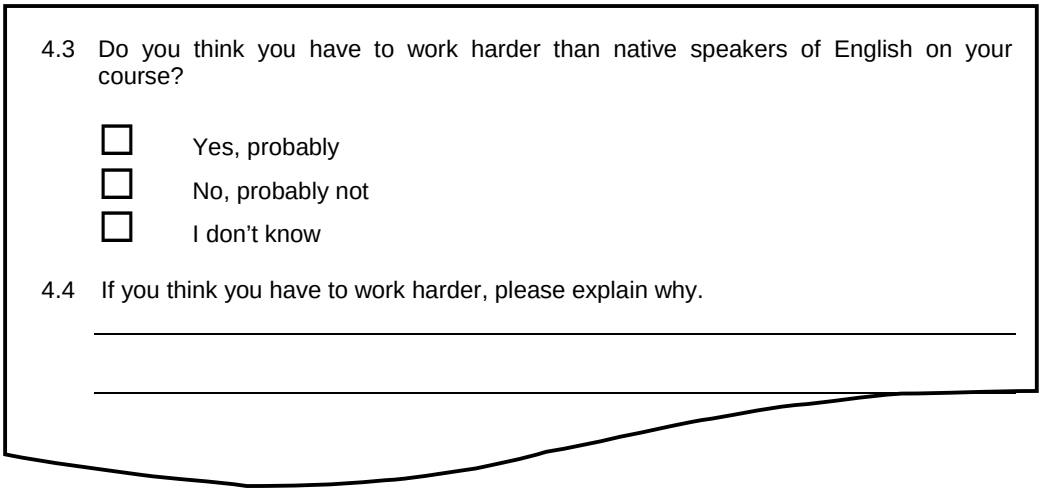


**Figure 6: Open and closed questions**
**Taken from Allwright & Banerjee (1997)**

The first question (4.3) is an example of a closed question. The respondent is asked to choose from one of three responses. Another common closed question type is one that uses a scale:

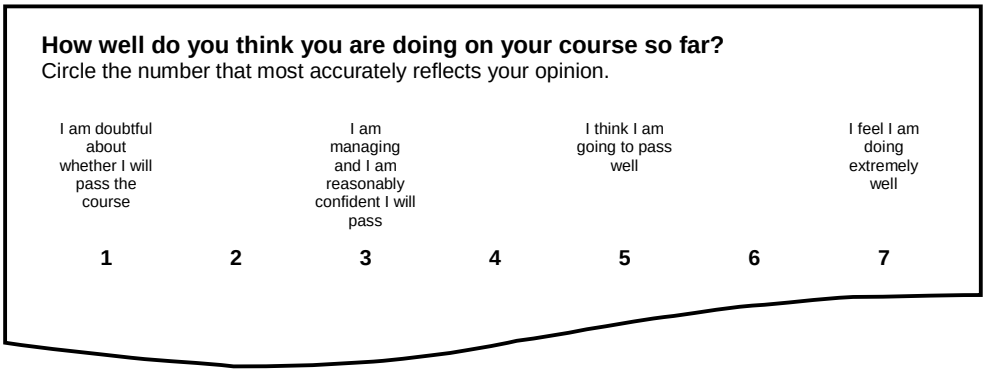


**Figure 7: An example of a questionnaire item using a Likert-scale**
**Taken from Allwright & Banerjee (1997)**

Note that only four points on the scale have been described. Some scales describe all the points and others describe only the two extreme points. You will need to decide how much guidance to give your respondents. It is important to bear in mind, however, that you cannot guarantee that your question will be clearer (and less open to interpretation) if you provide more guidance. Low (1996) has demonstrated the minefields within rating scale wording (e.g. Likert scales), pointing out a number of pitfalls, including:

1. describing the midpoint. If you offer your respondents a midpoint on the scale (e.g. '2' on a three point scale), you need to think carefully about whether the midpoint represents neutrality (i.e. neither agreement nor disagreement with the proposition) or undecidedness about the proposition (i.e. 'I don't know').
2. the number of dimensions that your options capture. Low (1996: 71) provides an interesting example where respondents have to say whether a course has helped them or not. However, the options that they can select include other dimensions such as enjoyment (e.g. 'I've had a lot of fun') and changes in proficiency (e.g. 'I've improved immensely').

The only way you can check that your questionnaire items are clear and are likely to be interpreted similarly by most respondents is by validating them (see section 7.6 for further discussion).

The follow-up question in figure 6 (4.4, above) is an example of an open question. Here, the respondent is asked to explain their answer and they can decide how much or how little they would like to say and what information they would like to provide.

Open questions can also be used on their own. For instance, in order to gauge attitudes to the IELTS test, questionnaires in the IELTS impact study (Banerjee, 1996; Horák, 1996 and Winetroube, 1997) asked both students and teachers to describe three things that they liked most about the IELTS test. Respondents were also separately asked to describe three things that they liked least about the test. Both these questions were deliberately open so that respondents could decide for themselves what they wished to include.

Open questions are particularly useful when you are not sure what the range of responses is likely to be (i.e. if your research is an initial exploration of issues) or if you want to avoid 'suggesting' answers to your respondents. You will find it easier to use closed questions when you are certain of the possible range of responses and/or when you want to make sure that you gather information on all the possibilities. In other words, you want to make sure that no possible response is accidentally forgotten.

It is important to note, however, that each question type has advantages and disadvantages. The advantages of closed questions are that they are quick to answer, process and compare. However, closed questions provide no scope for other answers and can reflect the researcher's bias in the categories provided. For instance, if you look more closely at the closed question presented in figure 6 you will see that the responses assume that the students' should compare their **overall** effort to that of their native-speaking classmates. However, further research by Banerjee (2003) has shown that students' experiences differ from subject to subject within a particular degree programme. For instance, MBA students with a background in Engineering find the more quantitative courses such as Management Science relatively easy. They find that they do not need to work harder than their native-speaking classmates on these courses. However, these students find that they struggle with less familiar and more language-oriented subjects such as Behaviour in Organisations. Therefore, the students will find it hard to give a single answer to the question 'do you think you have to work harder than native speakers of English on your course?'. Indeed, respondents could become frustrated or irritated if the response options did not suit what they wanted to say.

Open questions, on the other hand, provide more scope for a variety of answers and also allow the researcher to probe answers (e.g. 'please explain your answer'). But, such questions are time-consuming to complete and demand more effort and commitment from respondents. It is also more time-consuming and difficult to code and analyse the responses. In particular you will need to interpret responses in order,

for instance, to decide whether two differently worded answers from two respondents mean the same thing.

The foregoing discussion has revealed that open and closed questions are equally useful and both have drawbacks. Indeed, there is no perfect question type. Rather, you should select the best type for your purposes. In most cases, you will decide to use a combination of open and closed questions as this will allow you to combine focused and proscribed questioning with some more exploratory prompts. Regardless of the question type you select you also need to think carefully about the wording of your questionnaire. Check your draft questionnaire for the following pitfalls:

i. double-barrelled questions – your respondents are likely to find the question difficult to answer and you will find it impossible to determine whether the answer refers to only one (indeed which one) or both parts of the question.
ii. unclear instructions – so respondents are not sure what to do.
iii. questions that do not apply to the respondent – it is important to allow respondents to indicate when a particular item does not apply to them.
iv. questions that rely on memory or are hypothetical – e.g. the responses to such questions are unlikely to be stable or accurate.
v. biased options – respondents might be uncomfortable about selecting an option that has been presented in a negative light.

Beware also of mixing positively phrased items with negatively phrased ones. If your respondents do not read each question carefully, they might give the wrong response:

> I think it is important to check the dictionary when I do not understand a word
> I do not think it is important to check my work after I have finished writing

Oppenheim (1992) and Dörnyei (2003) provide good overviews of questionnaire design. Dörnyei (2003) gives particularly practical advice on the length and layout of the questionnaire. In particular he advises researchers to resist the temptation to include every question that they think might be useful. He warns that a questionnaire should not take more than 30 minutes to complete. He also reminds us that we need to take into account the reading speed of our respondents (2003: 17 – 18). Therefore, if you are gathering questionnaire data from young learners (e.g. 10 – 12 year olds) or are administering your questionnaire in a student's L2, then you need to consider how quickly they will be able to read and respond to the questions. Indeed, you should also to make your wording simple and accessible to the lowest level student you are gathering data from.

Dörnyei's (2003) advice makes it clear that questionnaire design is very complex and requires you to be very clear about the information you are trying to gather and also to think carefully about how to elicit that information in the most economical way possible. I would suggest the following six-step procedure for questionnaire design:

1. Brainstorm all the areas and possible questions that your questionnaire should cover.
2. Write questions to address each of these areas.
3. Return to the original purpose of your questionnaire. Eliminate all the questions that do not address that purpose.
4. Group the questions so that you can see where overlaps exist. Examine the overlaps in order to decide whether or not they are necessary. Bear in mind that you might want to ask the same question twice (in slightly different ways) in order to check the stability of your respondents' views.
5. Format the questionnaire and administer it to a small group of target respondents. Ask them to mark the questions that they do not understand. Time how long it takes for each respondent to complete the questionnaire.

6. Re-work the items that were difficult to understand. If the questionnaire was too long, consider carefully whether you can remove any questions without damaging the coverage of your questionnaire.

Questionnaires can be used to investigate test quality in a number of ways. For instance, they can be used to gather feedback from test-takers. Brown (1993) explored the usefulness of test-taker feedback questionnaires for the test development process. She gathered feedback from 53 test-takers during the trialling of a tape-mediated test of spoken Japanese for the tourism and hospitality industry – the Occupational Foreign Language Test. The questionnaire had two parts. In part one, the test-takers were asked for their overall attitudes to the test. For example they were asked if the test reflected accurately how well they spoke Japanese and whether they believed that the test reflected the type of language they would need in the tourism and hospitality industry. In part two, the test-takers were asked to comment on individual sections of the test. They were asked to rate each section for its usefulness and difficulty and also to say whether they had had enough time to respond. The test-takers were also encouraged to make comments on any items that they found problematic. Brown (1997) commented that the survey results confirmed that the content and level of the test was appropriate for the target language use situation. She also reported that the results revealed a lot about the expectations of the test-takers and indicated that much more advance information was needed. This feedback was used to improve the test handbook.

Clapham (1997) also used questionnaires during the test development process. She presented the revised IELTS test and specifications to different stakeholders, along with a detailed survey instrument that asked for their views on the extent to which the revised test sampled the test specifications. The questionnaires are presented in full in Clapham (1997: 133 – 140). One questionnaire was sent to academic subject specialists who would teach students who were admitted to university on the basis of their IELTS scores. This instrument included questions about whether the texts were comparable to the sorts of texts that students would have to read in their academic courses and whether the reading tasks were comparable to the reading tasks that students would have to perform on their academic courses. The subject specialists only had to look at one version of the IELTS test in order to answer these questions. The second questionnaire was sent to language teachers, testers and applied linguists. It contained the same questions as the questionnaire for subject specialists but the language specialists were asked to look at more versions of the IELTS test. The results of these questionnaires were used to make changes and improvements to the specifications of the IELTS test.

Marinič (2004) has demonstrated how they can be used during the test piloting phase to gather feedback from test-takers. She showed that test-takers can be asked for their views on the topics and methods of the test-tasks, the clarity of the instructions and also whether they were given sufficient time in which to complete the tasks. Additionally students can be asked whether they found the task difficult. In some studies, students have been asked to estimate whether or not they got the item correct as well. Marinič (2004) explained that this data could be analysed alongside the item statistics available for the tasks in order to judge the quality of individual test tasks.

Data can also be gathered routinely after live administrations. Halvari & Tarnanen (1997) described a study of the Finnish National Certificate language tests. The National Certificate tests can be taken in a number of different languages but the most commonly taken languages are Finnish, Swedish and English. It is not uncommon for a test-taker to sit for a test in more than one language. Such test-takers are a good source of information about the comparability of tests at the same level but in different languages. Halvari & Tarnanen (1997) distributed questionnaires after the test administration to test-takers who had taken tests in more than one language. The test-takers were asked whether they agreed with the scores that they had obtained (both the overall score and their score for each sub-test). They were also asked to identify any differences between the contents of the tests in the different languages. Halvari & Tarnanen (1997:

134) categorised the comments they received into three basic groups. They found that test-takers commented on:

1. differences in the test-taking context (e.g. one test-taker commented that the room for the German test was very cold).
2. the relationship between their test result and their 'true' language ability.
3. differences between the content of the tests (n.b. some of these were comments about test difficulty i.e. the English test was more difficult than the Swedish test).

Despite some interesting results, Halvari & Tarnanen (1997) found that the response to their questionnaire was rather low. This made it difficult for them to draw specific conclusions. Nevertheless, they argued that such data can throw light on the tests from the test-takers' perspective and can be used to make improvements to the test conditions and tasks.

Another use of questionnaires is to gather background information about test-takers. Test-takers routinely provide information when taking the IELTS through the Candidate Information Sheet (CIS). This instrument asks test-takers for their gender, age, language background and other language learning information. Herington (1997) developed a more detailed version of the CIS as part of the IELTS impact study project (described above). This questionnaire included questions about the students' attitudes to learning English and to taking English tests. They were presented with a list of statements about learning English and taking English test and asked to indicate how strongly they agreed or disagreed with each statement. For instance:

| | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| English is an easy language to learn | | | | | | | |
| I feel nervous when I see new words in an English test | | | | | | | |

Herington (1996: 48)

Here *–3* represented 'strongly disagree' and *3* represented 'strongly agree'.

Herington's (1996) instrument also asked test-takers to describe their learning strategies and their test-taking strategies. For instance:

| | 0 | 1 | 2 | 3 | ☺ |
|---|---|---|---|---|---|
| I learn new words in English by translating them into my language. | | | | | |
| During an English test the first thing I do when I read a passage is to look for the main ideas. | | | | | |

Herington (1996: 49 - 51)

The scale for this section ranged from 0 (never) to 3 (always). It also included an interesting additional option - ☺. This symbol meant 'a good idea but I don't do it'. Herington (1996) hoped that this would help test-takers to be very accurate in their claims about the strategies they used.

Background information questionnaires such as the one Herington (1996) designed can be used when analysing test-takers' performances on the test. The results can be categorised according to country of origin, language background and gender. Such analyses are routinely performed by testing organisations such as the Educational Testing Service (ETS – http://www.ets.org). You might even analyse the better (or worse) performers in more detail to see if they use common learning or test-taking strategies. This information can be used to give advice to future test-takers about how to prepare better for the test.

You might even wish to gather specific background information if you are considering major changes to your test. When ETS was preparing to introduce the Computer-based Test of English as a Foreign

Language (TOEFL CBT) they conducted a number of computer familiarity studies across the world (Kirsch et al., 1998; Eignor et al., 1998 and Taylor et al., 1998). In their first study they surveyed 90 000 test-takers. Each test-taker was asked to provide some background information such as their country of origin, educational background and language background. They were also asked to complete a computer familiarity scale. For instance, test-takers were asked how often they had access to a computer. They were also asked where they had access to a computer (e.g. at home, at work etc.). Test-takers were also asked to indicate how often they used the computer for specific tasks such as surfing the Internet. The responses to this familiarity scale were analysed to give profiles of the computer familiarity of test-takers in different parts of the world and from different backgrounds. A second study was then carried out to compare the test-takers familiarity with computers to their performance on a set of 60 computer-based TOEFL tasks. Each test-taker first took a computer familiarisation tutorial which trained them in the computer skills that they would need in order to take the computer-based TOEFL (e.g. how to use a mouse). Taylor et al. (1998) report that there was no evidence that the computer delivery of test items affected test-taker performance (regardless of the test-taker's previous computer familiarity). This indicated that the familiarisation tutorial provided sufficient support to test-takers who were unfamiliar with computers.

Other test-taker characteristics might also affect the construct validity of a test. For instance, Allan (1992) developed a scale of test-wiseness in order to explore the effect of test-taking strategies upon test-takers' performance on a reading test. He argued that the test-taking skills of L2 learners had little to do with their reading abilities yet could affect their final reading test scores. Allan (1992) developed a 33-item instrument and administered it to 51 students in a Hong Kong polytechnic. Each item was a multiple-choice question. The test-takers had to answer the question by choosing the most appropriate option from the choices. The items were designed such that the test-takers would not be able to answer them from their background knowledge. The correct answer was 'cued' in one of the following ways:

1. stem-option (there is an association between a word in the stem and a word in one of the alternatives. This association is usually semantic or grammatical).
2. grammatical cue (the option grammatically matches the stem e.g. the form of the article might suggest that the option should begin with a vowel sound)
3. similar option (this is when all but one of the options are similar in meaning. This makes the 'odd' option stand out)
4. item giveaway (the answer to the item can be found in another item)

Approximately one third of the students were also asked to provide brief explanations for their answers. This data was used to throw light upon the responses. Allan (1992) found that the items in the 'grammatical cue' and 'item giveaway' sets appeared to correlate well with one another. The results for the other two sets ('stem-option' and 'similar option') were less clear. Nevertheless, he argued that some students were more sophisticated test-takers. He further suggested (1992: 109) that this was particularly problematic for teacher-designed tests because these were less likely to be carefully piloted and validated.

Questionnaires can also be used to investigate the processes used by test-takers to complete different items. Li (1992) administered a questionnaire within a reading test in order to explore which reading strategies each test-taker used to complete individual items. The test-takers first completed an item and then indicated which of a list of reading processes they had used to do the item. He also asked them to indicate whether they found the item difficult or easy. Li's (1992) analysis of the questionnaires confirmed the findings of Alderson (1990) that test-takers use a variety of reading skills to complete test items. While some overlap may exist, in general it is very difficult to predict the reading skills that test-takers will use to complete a particular test item. This research cast doubt on whether test constructors can design items that test specific skills.

The studies described in this section have shown that questionnaires can be used in a number of ways to examine test quality:

1. To canvas test-taker views on the difficulty and/or appropriacy of test items.
2. To explore the views of other stakeholders such as teachers, test designers and applied linguists on the suitability of test input and test tasks for the target group of test-takers.
3. To gather information about test-takers in order to profile the test-taking population.
4. To establish the need for test-taker training or familiarisation as well as the nature of that training.
5. To investigate possible threats to construct validity (such as the influence of test-wiseness or computer familiarity upon test-taker performance).
6. To explore test-taking processes and strategies.

Questionnaires can also be used at various stages in the test development process as well as during live administrations. It is important to note, however that questionnaire response rates can be low. Indeed, Halvari & Tarnanen (1997) report that only 63% of the questionnaires they distributed were returned and return rates can sometimes be as low as 30%. It is therefore better to ask respondents to complete questionnaires in your presence (either in class or immediately before test-takers are released from the testing venue). This ensures that they have to hand in the questionnaire before they leave.

### 5.2 Checklists

If you have ever taken a car for a routine service you will probably have noticed that the mechanic has a form that must be filled during the procedure. The form comprises a list of features that must be checked. The mechanic is required to tick every item off and also to note any problems in a space provided. This is a checklist.

Checklists are used in a variety of contexts including store inventories and quality control inspections. They are also very useful in investigations of test quality. The key feature of checklists is that they structure observations. As such they can vary in format from very clearly defined lists, where the researcher simply ticks for the presence or absence of a particular feature or characteristic, to more open grids. In their more open form, checklists might simply comprise a list of column or row headings with space in which to make notes. The checklist for validating speaking tasks developed by O'Sullivan et al. (2002) falls into the former category, while the Communicative Orientation of Language Teaching (COLT) observation instrument developed by Allen et al. (1984) falls into the latter category. Alternatively, a checklist might combine elements of the two as does the Classroom Observation Instrument designed for the IELTS impact study project (Banerjee, 1996). The first three pages of this instrument comprised an open grid that asked observers to note the time taken for each activity, what the teacher did, what the students did and the nature of the interaction. The remaining pages listed different task types and text types as well as different interaction patterns. The observer was asked simply to tick the task types, text types and interaction patterns that he/she observed.

It is rare for a checklist to be adopted directly from another context. Instead, researchers usually survey and analyse other checklists, paying attention to the features that might be useful in their context. Banerjee (1996) used this process when she designed the Classroom Observation Instrument for the IELTS impact study project. She first analysed the COLT observation instrument (Allen et al., 1984) and an instrument designed for the Sri Lankan impact study (Wall & Alderson, 1993). These proved very useful in suggesting an overall design for the observation instrument. In order to identify specific items to include in the checklist, Banerjee (1996) needed to decide what washback from the IELTS might look like. To achieve this she closely examined the test materials and published teaching materials available for the test (in this case the IELTS test). She also brainstormed the content of the checklist with other researchers, teachers and students. Though this was not possible in the case of the IELTS impact study project, it is also advisable to analyse the test specifications. Additionally, it is always useful to attend a typical lesson in order to document the teaching and learning that takes place (either through field notes or a video-

recording). This will enable you to identify categories of data that you would like to capture. All these sources of information (test materials, specifications, published teaching materials, brainstorming etc) will help you to compile a full list of the activities, interactions, text-types etc. that could occur in a typical lesson. A detailed checklist can then be produced.

The checklist should then be extensively trialled and revised until you are sure that it is easy to use and will also help the observer to capture all the information being sought. Banerjee's (1996) observation checklist was reviewed by the Language Testing Research Group at Lancaster University, a group of researchers, teachers and students with a lot of experience in designing research instruments. Banerjee (1996) also trialled her observation checklist with an IELTS-type class in order to ensure that it was practical to use in a live observation. She conducted this observation exercise with a colleague with whom she was later able to compare notes. This comparing of observation notes revealed the extent to which the observation checklist helped the two observers to make note of the same features of the lesson (a reliability check).

As has already been stated (above) Banerjee's (1996) final instrument combined an observation sheet and a checklist of activities, interactions and text-types. It was very similar in structure to the observation checklist that Wall & Alderson (1993) used when they investigated the effect of the introduction of a new Secondary School leaving test ('O' level) upon the teaching that took place in Sri Lankan classrooms. At the time little empirical research had been carried out to establish the influence of a test upon teaching and learning in the language classroom. Wall & Alderson's (1993) study was also innovative in that it included direct observation of classrooms whereas previous research had been based on questionnaires and interviews. Indeed, it is important to note that the data gathered from questionnaires and interviews is self-report data i.e. what teachers, students and test-takers 'say' they do or believe. It is often useful to complement such data with direct observation such as classroom observation or the observation of live test administrations in order to, as Wall & Alderson (1993: 42) argue, take into account not only what study participants report about the effect of an exam upon their teaching, learning and/or test-taking practices, but also to capture what those practices might look like in reality.

Wall & Alderson (1993) hoped to examine the extent to which the new Sri Lankan English 'O' level had influenced the types of teaching activities that took place as well as the interaction patterns (e.g. teacher-student or student-student interaction) and the input text types. Therefore, their observation instrument included checklists of different teaching activities, interactions and input text types. These lists included activities, interactions and text-types that occurred in the test as well as other activities, interactions and text types that were not represented in the test and which it was hoped would not occur in the classroom because they were considered to be poor teaching practice. A copy of this observation checklist can be found in Alderson & Wall (1992).

The observations were conducted by seven Sri-Lankan teachers, each of whom visited 7 schools six times over a period of two years. It is important to note that the six rounds of observation were carefully timed to capture different 'moments' in the academic year. For instance, round 1 took place at the start of the first year, round 2 was scheduled for the middle of the school year (four months after the first observation round and three months before the examination). Round 3 took place shortly before the examination. Rounds 4 – 6 followed the same pattern in the following academic year. Wall & Alderson (1993) encountered a number of difficulties in the data-gathering for this study. Firstly, the round 1 observations were disrupted by political instability in Sri Lanka. The round 3 observations were also affected, this time by the fact that students were released from regular classes more than one month before the examination so that they could study for the exams. Wall & Alderson (1993) also had to cope with changes in the team over the two-year period of the study. Finally, the Sri-Lankan teacher-observers sometimes had difficulty in being released from their regular teaching duties in order to conduct the observations. These difficulties are instructive because they are not unusual. Any study will have to take into account the 'rhythm' of the

teaching year (including the fact that teaching might be suspended early for examination classes) as well as the availability of research participants and helpers. It is always important to gain the support of governing bodies so that you can maximise the co-operation you might expect for your study.

Despite the difficulties they encountered Wall & Alderson (1993) reported that they had a full data set (i.e. 6 rounds of observation) for 18 schools. Also, at its largest the sample contained 64 schools (the observations from round 5). Even though the smallest round of observations contained data from only 18 schools, the second smallest round of observations included a creditable 36 schools. Most of the data that was gathered through the observations was analysed using the statistical software tool SPSS (http://www.spss.com) to calculate the frequency of occurrence of particular features. For instance, Wall & Alderson (1993) calculated the percentage of classes that were devoted to the different language skills (reading, writing, listening, speaking and language form). This amounted to a quantitative analysis of data that had been collected using a qualitative data collection method. This is not unusual for the analysis of questionnaires and checklists. Indeed, quantitative analysis of data is a useful complement to qualitative analyses and Wall & Alderson (1993: 55 - 57) also looked carefully at patterns in the teaching methodology, reporting a tendency towards a lockstep approach where the teacher dominated the interaction. As a result of this combination of analyses, Wall & Alderson (1993: 66) reported that the Sri Lankan 'O' Level examination had some effect on the content of teaching and upon the design of in-class tests in Sri Lankan classrooms. However, they could not find evidence of the effect of the examination upon the method of teaching.

A recent and rather different example of a checklist is the observation checklist developed by O'Sullivan et al. (2002) to validate speaking tasks. O'Sullivan et al. (2002) were motivated by the fact that most speaking test validation requires detailed and time-consuming analyses of test language as has been described in section 3 (above). They wanted to develop a framework that could be used during live administrations to analyse the language elicitation tasks (LETs). They argued that the performances elicited by LETs should match the predictions of test designers if we are to make valid interpretations of test-takers' scores but also contended that analyses of test language (the most common method for analysing speaking test performances) were time consuming and demanded considerable expertise. Consequently, the sample of test performances subjected to such analyses tended to be small and was therefore not easily generalisable. O'Sullivan et al. (2002: 39) argued for a methodology that complemented more detailed analyses of language samples but could be applied to larger numbers of test takers.

O'Sullivan et al. (2002) began by reviewing the literature in spoken language, second language acquisition and language testing in order to identify a set of informational and interactional functions that can occur in spoken language. Three lists were written initially and these were then refined via a number of meetings in which participants used the checklists and then commented on their usability. Through this process, items on the checklist that could not achieve a high degree of agreement were discarded and other items were improved to make them more transparent. The final version of the checklist is presented in O'Sullivan et al. (2002: 54). It consists of three categories of functions: informational functions (including providing personal information, speculating and describing), interactional functions (including agreeing, modifying and asking for information) and managing interaction functions (including initiating, reciprocating and deciding). This checklist is a good example of the way in which a data collection framework can be developed and used in post-hoc analyses of test output. It is important to note, however, that the final form of the checklist was influenced by the Cambridge ESOL tests on which it would be used. This is further evidence of my earlier claim that observation instruments like checklists are rarely adopted directly from another context. They are more likely to be customised to the test being investigated.

The research reported so far has demonstrated that checklists can be used to investigate test quality in the following ways:

1. To explore the impact/washback of a test upon the teaching and learning in the language classroom.
2. To investigate the match between test-designers predictions and the actual language elicited by test tasks.

Checklists can also be used during item moderation meetings. Observers can use them to record the decisions that are taken with respect to individual items and the test as a whole. Similarly, checklists can be used during rating scale development. The resulting data can reveal a great deal about the construct of the test as well as the thought processes of item writers and test and scale developers. Additionally, test-takers can be observed while they are taking the test and assessors can be observed during the rating process (as a complementary procedure to verbal reports). It is clear, however, that checklists are used in these contexts to structure observation. Finally, you will probably also find it useful to audio or video record events such as item moderation meetings and assessor moderation exercises. The transcripts from these recordings can later be analysed in greater detail.

The studies reported here also indicate that checklists (like questionnaires) can be used to collect larger samples of data in a systematic and easily comparable manner. However, there are also some key considerations:

1. Stability of the group that conducts the observations. Wall & Alderson (1993) found that their observation team changed from one study year to the next. Additionally, their observers were also teachers and sometimes found it difficult to get leave from their teaching responsibilities in order to carry out the observations.
2. Training for the observers. O'Sullivan et al. (2002: 46) argue that observers should be trained to use the checklists "if a reliable and consistent outcome is to be expected". As with the use of task characteristics frameworks (see section 4.1), training will inevitably result in 'cloning' of observers. However, this is important if you intend to compare and combine different observations.
3. Observation checklists should be piloted extensively and validated carefully to ensure that they are performing appropriately in the context for which they are used. Validation issues will be discussed in section 7.6 (below).

### 5.3 Interviews

The final feedback method to be discussed is the interview. It is probably best described as "a conversation between interviewer and respondent with the purpose of eliciting certain information from the respondent" (Moser and Kalton, 1971: 271) and has many of the same purposes as questionnaires. It differs from questionnaires primarily because it is a more flexible data collection method; a questionnaire item is pre-prepared and cannot be altered at the point of administration whereas an interview question can be altered to suit the flow of the interaction between the interviewer and the respondent. Yet questionnaires and interviews should not be viewed as polar alternatives. You will probably find that they combine well with each other. Questionnaires can be used to gather information on a set of clearly defined themes from a large number of respondents (some sample sizes exceed 1000 respondents) while interviews can be used to probe some themes in greater depth and detail with a sub-set of the questionnaire respondents.

Interviews can take a number of different forms. They can be individual (where there is one respondent and one interviewer) or group (where there are two or more respondents and one interviewer) interviews. Individual interviews have the advantage of your being able to focus in considerable detail upon the views of a single respondent and to build a picture of an individual test-taker or stakeholder. However, group interviews can be used to brainstorm ideas and to establish group viewpoints. One advantage of the group

interview is that the interaction between respondents can sometimes spark revelations that you, as the interviewer, might not succeed in eliciting from a single respondent.

Interviews can also vary in their degree of structure. Regular census data is often collected by structured interview. The interviewer either contacts you by telephone or by coming to your front door. He/she has a fixed schedule of questions to ask. The wording and the order of the questions is pre-determined. At their most structured, such interviews closely resemble questionnaires. Shohamy et al. (1996) conducted structured interviews with teachers and inspectors as part of their investigation into the impact of two national tests - an Arabic as a second language test and an English as a foreign language test. The interviews included questions about preparation for the test, stakeholders' knowledge about the test and the impact of the test upon teaching and testing practices (1996: 302). This data was complemented by data from questionnaires administered to students and an analysis of test documentation such as bulletins issued by the Ministry of Education.

Unstructured interviews fall at the opposite end of the continuum. The ground covered in these interviews is dependent upon the interaction between the respondent and the interviewer. The latter rarely has more than a set of themes to guide the discussion. Though this is the most flexible of the interview structures, it is also the most demanding. If poorly handled, interviewers risk that the interview data will not result in helpful or interesting revelations. Indeed, such interviews are usually best conducted by highly experienced and well-prepared interviewers.

Taking the middle ground are semi-structured interviews where the interviewer has an interview schedule to guide the discussion but where there is some room for the respondent to negotiate the pace and coverage of the interview. Allwright & Banerjee (1997) used this type of interview in their investigation of the study experiences of non-English speaking post-graduate students at a British university. They selected this interview type for a number of reasons:

1. They were each going to interview half the students in a series of individual interviews. Consequently, they needed to have a structure to follow so that their respective interviews yielded comparable data.
2. Though their concern for having comparable data suggested the use of a structured interview, Allwright & Banerjee (1997) wanted to retain some flexibility to respond to the themes that emerged during the interviews.

Since the semi-structured and unstructured interview allow the interviewer to respond to the data as it emerges, this also means that these interview types have a distinct social dimension. Consequently, their direction and success can be influenced by the interaction between the interviewer and the interviewee. Banerjee (1999) compared the interviews she conducted as part of the Allwright & Banerjee (1997) study with those conducted by Joan Allwright (the lead researcher on the project). Banerjee's (1999) analysis of the transcripts revealed that the interviews between herself and the study respondents were slightly strained in comparison to those conducted by Joan Allwright. The students she interviewed appeared unwilling to respond to questions that probed their responses. In contrast, the students interviewed by Joan Allwright seemed generally more willing to elaborate and often stayed well beyond the agreed time limit for the interview. Banerjee (1999) viewed this experience as an example of what Mishler (1986) describes as the co-construction of the interview by the participants. She argued that the interviews were different because the people involved were different and the interpersonal dynamic therefore differed. She contended further that the key to that different dynamic lay in the relationship she had with the respondents compared to Joan Allwright's relationship with them. At the time she was the research assistant on the project and a research student. As such she was the respondents' equal – a fellow student. In contrast, Joan Allwright was a member of staff. Banerjee (1999) argued that this power differential at least partly determined the tendency of the respondents to be more forthcoming with Joan Allwright and

less impatient to end the interview. They possibly wanted to appear co-operative for the interviewer they perceived to be in a superior position to them.

It is, of course, also possible that one interviewer may be more experienced and therefore more skilled than the other. This underscores the importance of preparing thoroughly for interviews. Borg & Gall (1983) advise that it is important to eliminate any bias that might be introduced by factors such as the length and location of the interview, the attitude of the informant to being interviewed and/or to the researcher and the behaviour of the researcher. It is clear, therefore, that interviews should be designed and piloted carefully. Always ensure that the interviewer has had an opportunity to practice conducting interviews before he/she begins collecting data. Indeed, if you plan to use a team of interviewers, it is useful to conduct an interviewer training session in which each interviewer can practice his/her interview technique as well as analyse and reflect upon the practice interview. If combined with a piloting procedure, the interviewer training can be used to refine and clarify the aims of the interview for all the interviewers.

It is important to note, however, that training and piloting will not eliminate (or render inconsequential) the effect of the interpersonal dynamic between interviewer and respondent upon the interview. I would recommend that, where possible, you should try to include familiarisation questions that allow the interviewer and respondent to relax in one another's company. You will probably also find it useful if you systematically note details about the interview situation such as the place, physical setting (arrangement of furniture, position of participants relative to one another) and the relationship between the interviewer and the respondent. This is because, as Stimson (1986) argues, data analysis should take account of the effect that the data collection setting might have upon the respondent.

As I have already said, interviews are rarely the only data collection method in a study. They tend to be combined with at least one other method such as observations (e.g. Alderson & Hamp-Lyons, 1996) or questionnaires (e.g. Shohamy et al., 1996 and Allwright & Banerjee, 1997). They are useful in investigations of test quality because stakeholders (including test-takers, teachers, administrators and parents) can be asked for their views about the test including the overall quality of the test (the extent to which they believe the test gives a true picture of language ability), the difficulty of specific tasks, items or input texts and the extent to which the input texts and tasks are interesting and/or authentic. Interviews can also be used to examine how test scores are interpreted and used by receiving institutions and other stakeholders.

Clearly, the advantage of interviews is that the interviewer can concentrate on a single respondent and thoroughly explore his/her views on the test. The interviewer can also probe responses in order to better understand the respondents' views. In this way interviews can provide a wealth of detail that might not be available from a questionnaire. However, interviews can be time-consuming (an interview can take an hour or more to complete). This means that fewer informants can be studied, which can in turn affect the generalisability of your results.

## 6. Using qualitative methods for standard-setting

I suggested at the start of this chapter that the qualitative methods described here could also be used for standard-setting. You will have read in the chapter on standard-setting (see Section B) that the establishment of cut-off scores involves expert judgements. You will also know that it is important to safeguard the validity of these judgements. This can be done using qualitative procedures. This area of research is still rather new so there is little published guidance on how to use qualitative methods to establish the validity of standard-setting procedures. This sub-section will suggest a few procedures that could be applied during the judgement phase (when standards are set) as well as during the specification phase (when the content coverage of the test is examined).

During the judgement phase it is necessary to establish benchmark performances for the productive skills (writing and speaking) and to establish benchmark texts, items and responses for the receptive skills (reading and listening) as well as for tests of linguistic competence (e.g. grammar and vocabulary) (for more details see Chapter 5 of the Manual). Expert judges establish these benchmarks by placing texts, items, responses and/or performances in the CEF bands A1 – C2. This process can be monitored and investigated as follows:

1. Judges can be asked to prepare their assessments individually. A meeting can then be convened in which each judgement is discussed.
2. The discussion can be recorded and observation notes can be taken.
3. The observation data and the transcripts of the recordings can be analysed later to explore the sources of agreement and disagreement more closely. This will throw light on the characteristics of test items, input texts, test-taker responses and/or performances that signal a particular benchmark. It will also help to explain the features of test items, input texts, test-taker responses and/or performances that can cause variation in expert judgements.
4. Additionally, selected participants could be asked to perform a retrospective verbal protocol. It might be helpful to use a stimulated recall protocol if the verbal protocol takes place a few days or weeks after the benchmarking meeting. This data could explain how the judges made their benchmarking decisions. It might reveal criteria unrelated to the performance or test input that have influenced the benchmarking decision. The latter could constitute a threat to the validity of the benchmarking.

This data could also be used to establish the validity of the final benchmarks and could inform future training and familiarisation programmes for expert judges.

Cut-scores are also estimated during the judgement phase. Subsequently, test-takers who receive scores above the cut-score will be presumed to have met a particular performance standard. Test-takers whose scores fall below that cut-score will be presumed not to have fulfilled the requirements for that standard. Yet, as Kaftandjieva (Section B of this volume) points out, cut-scores are arbitrary. It is necessary, therefore, to gather evidence of the validity of the final cut-scores in order to legitimise them. But the validation of standards is not achieved by an appeal to external criterion (Kane, 2001). Instead it is important to gather evidence to support the cut-score decision. This can be done by demonstrating that the decision-making process was logical and reasonable and that the decision is plausible. Qualitative evidence could be gathered at the following points in the process of setting a cut-score:

1. The meeting at which individual judges discuss their individual conclusions about the cut-score can be recorded and observation notes can be taken. The observation data and the transcripts of the recordings can be analysed later to explore the sources of agreement and disagreement more closely. This will throw light on the characteristics of test-taker responses that signal a particular level of performance. It will also help to explain the features of test-taker responses that can cause variation in expert judgements.
2. The transcripts and observation notes can also be analysed to demonstrate that the cut-score procedure was carried out correctly and with appropriate attention to detail.
3. It might also be useful to conduct follow-up interviews with the judges. The interview questions should ask for their views on the cut-score procedure. The judges should also be asked if they believe the final cut-score was appropriate and whether they felt able to be honest in their judgements during the setting of the cut-score. These interviews will provide evidence of the credibility of the procedure followed and also of the extent to which the final judgement is plausible.
4. Additionally, selected participants could be asked to perform a retrospective verbal protocol or a stimulated recall protocol of their own judgement process. This data could explain how the judges made their cut-score decisions. It might reveal criteria unrelated to the performance or test input

that have influenced the cut-score decision. The latter could constitute a threat to the validity of the cut-score.

During the specification phase judges are likely to be asked to examine the content coverage of the test. The judges will examine each input text and item to answer a number of questions such as:

i. Which situations, content categories, domains are the test takers expected to show ability in?
ii. Which communication themes are the test takers expected to be able to handle?
iii. Which communicative tasks are the test takers expected to be able to handle?
iv. What kind of communicative activities and strategies are the test takers expected to be able to handle?

(examples taken from Form A10, Council of Europe, 2003: 43)

The validity of this process can be established in similar ways to those described for the judgement phase:

1. The exemplar judgement sheet provided in the Manual, Form A10 (Council of Europe, 2003: 43), requires judges to provide evidence for their judgements. This evidence could be compared across judges to identify similarities and differences in the evidence selected to justify the judgements made.
2. A few judges could be asked to perform a retrospective verbal protocol or a stimulated recall protocol of their own judgement process. This data could explain how the judges performed the analyses and selected their supporting evidence. It might also provide additional insight into the judgement process that the judges had not written down.
3. It might also be useful to conduct follow-up interviews with the judges to explore the evidence provided in more detail. For instance, judges could be presented with the evidence that they did not provide and asked to discuss the suitability of that evidence. This will explain differences in the evidence provided.

The verbal protocol and interview data may also provide you with feedback on the usability of the forms.

## 7. General issues arising

The discussion so far has revealed that qualitative methods of investigating test quality share a number of theoretical and practical concerns. The more practical issues include deciding what language to collect the data in, how to go about piloting and trialling the instruments and what level of detail to provide in transcriptions. The more theoretical issues include decisions about triangulating data sources, analysing the data, the validity of the instruments and procedures and the generalisability of the results. In this section I will return briefly to each of these issues.

### 7.1 Language that the data is collected in

I commented in 2.1 that, when collecting qualitative data, the choice of language is not necessarily straightforward. It is relatively common for diary studies, interviews and questionnaires to be conducted in the respondents' L1 but the language of verbal reports has varied from study to study. Key issues to consider are:

1. The respondents' L2 proficiency. If you are gathering data from respondents with low language proficiency you might find it more productive to gather the data in their L1. This will enable you to probe for more sophisticated answers. Indeed, if you conducted the interview or verbal report in the respondents' L2 you might worry that the depth of responses was adversely affected by the respondents' L2 proficiency (regardless of their level of ability in their L2).
2. Your own ability in the respondents' L1. There are contexts in which the researcher does not speak the respondents' L1 well enough or at all. This could be because the researcher has not learned that language sufficiently well to conduct interviews or verbal protocol procedures with study participants. In such circumstances you might wish to work with a native speaker of the respondents' L1 who could gather the data on your behalf. However, this might not be a practical solution in cases where the study participants come from a wide variety of language backgrounds.

For example, in Allwright & Banerjee's (1997) study the 38 respondents represented 20 different nationalities, and spoke a range of 13 different languages. It would have been impractical to arrange for these respondents to receive questionnaires in their L1 and to be interviewed in their L1. Indeed, this might have further complicated the interpersonal considerations that arose with using two interviewers working separately to gather the data (see 5.3, above, for more discussion).

3. The cognitive load of performing a task in the L2 but talking about it in the respondents' L1 might affect the processes that you are trying to capture. In such circumstances, you might wish to gather the data in the L2 so that this cognitive load is controlled.

## 7.2 Piloting and trialling

It is important to pilot all the instruments that you use and to train everyone who will be involved in collecting data. Piloting of instrumentation is particularly important when you are gathering data using feedback methods such as questionnaires, observation checklists and interviews. Piloting is usually on a smaller scale than the main data collection phase but must be conducted with a comparable context and with a similar sample group of respondents. The purpose of the piloting stage is to check that the questions or observation prompts are eliciting the data that you are trying to capture and that your respondents understand the wording of the questions. Piloting also gives you feedback on your procedures for gathering the data. For instance, you can use piloting to establish the best time to administer a questionnaire or to check that your instructions and procedures are clear and efficient.

Observer- and interviewer-training is also important for successful data collection. Though the training phase could be combined with the piloting phase it is probably best to conduct observer and interviewer training after the instruments are finalised. As with piloting, training must be conducted in a comparable context to the live data collection context. In the case of observer training the data used can be pre-recorded. Observers can be asked to complete the observation checklist while watching a video recording of a class, test performance or test administration. They can then discuss the notes they have taken, using the video-tape record to discuss the aspects of the lesson, test performance or test administration that they did not capture. This discussion should alert the observers to aspects of the observation context that they should pay particular attention to. It should also familiarise them with the observation instrument. This process can be repeated until you and the observers are confident that they are ready for live data collection.

Interviewer training is rather more complex. Though video-recordings are useful for familiarising interviewers with the interview structure and alerting them to possible pitfalls, it is also important to give interviewers one or two practice interviews. Each practice interview should be recorded so that they can be reviewed. The practice should help the interviewers to internalise the interview structure and should help them to conduct the interview more naturally (with less recourse to notes). The discussion should alert the interviewers to possible pitfalls in their own interviewing style.

## 7.3 Transcribing the data

If you intend to analyse your data using Conversation analysis you will need to adopt the detailed transcription scheme that I described in 3.1. For other types of analysis, however, you need to pick the most appropriate level of detail for your purposes (Silverman, 1993: 124). Silverman also advises that you should adopt transcription conventions that are achievable within your constraints of time and resources (1993: 124). For instance, in her study of the influence of different language proficiency levels upon students' experiences on academic courses, Banerjee (2003) was primarily interested in **what** her respondents said about their study experiences rather than in the nature of the interaction between herself and her research participants. Consequently, she adopted a very simple transcription scheme for her interview data:

| | |
|---|---|
| **,** | pause for breath during a thought. |
| **.** | pause at the end of a thought. |
| **?** or **(?)** | a question either to self or to other speaker. |
| **!** or **(!)** | particular emphasis placed during utterance. |
| **mmm** or **um** | sounds usually indicating thinking. |
| **mhmm** | sound indicating agreement. |
| **…** | pause of any length. |
| **[unclear]** | speech that could not be decoded. |
| **[ ]** | action/event occurring or co-occurring e.g. [laughs] = laughter from speaker; [tape ends] = end of side A or recording. Also used for my own clarifications of what is being referred to e.g. [1998/199 class] clarifies which MBA class the speaker is referring to when she says 'class'. |

(Banerjee, 2003: Appendix 5J)

Banerjee (2003) captured repetitions and fillers (such as 'you know') but did not need, for her purposes, to capture the pace of delivery or pronunciation of her interview respondents. Similarly, she did not attempt to capture overlapping speech as this was not relevant to her analysis. Instead, she used standard punctuation (e.g. commas and full stops) to indicate natural pauses in delivery. However, she felt that non-verbal behaviour (e.g. laughter or a pause to check or read from a file) was relevant to her analysis so this was noted. Banerjee (2003) developed this transcription scheme iteratively while simultaneously analysing a subset of her data. This helped her to develop a transcription scheme with an appropriate level of detail.

It is important to note, however that you may not need to transcribe all (or perhaps any) of your data. In some cases it may be enough to listen to the recordings several times, taking detailed notes and transcribing only the most illuminating or colourful extracts. You can then report the broad themes thrown up by the analysis, flavouring it with appropriate extracts.

### 7.4 Triangulation of data sources

The perennial question that needs to be answered in any study is whether the data that was gathered was a true reflection of the reality it was intended to study. The triangulation of data sources refers to the gathering of data about a particular event or context from a number of different angles. If the data gathered from each of these perspectives or angles all suggests the same interpretation or conclusions, this can help to corroborate your claims.

Triangulation can be achieved in a number of ways. First, you could use two or more methods to collect your data from your respondents. For example, in their study of the effect of the TOEFL test on teaching Alderson & Hamp Lyons (1996) first interviewed the teachers and then followed this up by observing the teachers in both TOEFL-preparation and non-preparation classes. Another way of triangulating your data is to collect data from more than one source. For instance, if you were exploring the appropriacy of the content of a test you might ask three different groups to provide judgements – test developers, teachers and test-takers.

In addition to corroborating your analysis, triangulation provides opportunities for probing certain aspects of your data in more depth such as when you follow up a questionnaire with in-depth interviews with a sub-set of your sample.

## 7.5 Analysing the data

Arguably, good analysis begins with the appropriate and accurate storage and transcription of data. Dey (1993: 74) argues that "[g]ood analysis requires efficient management of one's data". It is important, therefore, that data is stored in a format that allows you to search it easily and to compare different transcripts. This can be done manually by using a system of filing cards and annotated transcripts. You might begin by highlighting and annotating your transcripts with themes and codes. Quotations could be transferred onto a filing card and labelled with the theme that they represent. If a quotation represents more than one theme, you could either complete two filing cards (one for each theme) or you could devise a cross-referencing system.

The manual approach is easy to transport but clearly very labour intensive and could involve a lot of repetitive work. Therefore, researchers are increasingly using electronic tools. There are a number of software packages that support qualitative data analysis, some of which interface with statistical tools like SPSS (see 5.2, above). Two examples of these are Atlas-ti (http://www.atlasti.de) and QSR NUD*IST (http://www.qsrinternational.com/products/ productoverview/product_overview.htm). These programmes help researchers to apply multiple codes to their data and to build theories about how the codes might be related to one another.

Nevertheless, data analysis tools cannot actually perform the analyses. They simply support the analysis being done. This phase of the research process can be very daunting for, as Denzin argues, data analysis "is a complex, reflexive process" (1998: 316) that involves making sense of the data and then representing it in a coherent way that explains the interpretation taken. The first question that must be addressed, however, is how to approach the coding. Indeed, the assembled data can be very overwhelming (cf. Buck, 1994 and Feldman, 1995). It is important, therefore, to find a way into the data perhaps by first looking for answers to your initial research questions or by inspecting your data for themes that have emerged from your review of the literature. For instance, Buck (1994) used his initial research hypotheses as his starting point when analysing his data. Another alternative would be to adopt the Grounded Theory approach (Strauss & Corbin, 1998). Grounded theory refers to theory that is data driven. It demands that researchers should look for patterns in the data rather than attempting to impose a pre-existing theory or explanation.

Regardless of the approach you adopt, however, Brown & Rodgers (2002) emphasise the importance of coding data in a way that helps you to reveal its underlying patterns. While the coding categories that emerge are usually specific to the research being conducted (e.g. Alderson (1990) coded for reading processes), Brown & Rodgers suggest three important considerations:

i. Are the code-categories clear and unambiguous?
ii. Is the coding scheme reliable? Will alternative analysts code data in the same way?
iii. Do the results of coding lead to useful analyses?

(2002: 64)

## 7.6 Validity, reliability, generalisability

This focus of this chapter has been on validity and how to establish that a test is valid. It follows, therefore, that the methods used to establish test validity should themselves be valid. As the Manual argues, "[i]n an empirical validation, the data have to be analysed and interpreted thoughtfully and with full awareness of possible sources of uncertainty and error" (Council of Europe, 2003: 99). Indeed, Maxwell (1992: 279) warns that the legitimacy of qualitative research is threatened when it cannot consistently produce valid results. Indeed, this is true of all research but the problem is perhaps more acute for qualitative research because of its interpretive nature.

Alderson & Banerjee (2001) provide a practical approach to instrument validation. Drawing on the procedures already used in test validation, they suggest a number of simple measures that can reveal the transparency and clarity of the language used in the instrument as well as whether the options offered (e.g.

never, sometimes, often) mean the same thing to different users and whether there are any gaps in the instrument. These measures include:

1. Reliability measures such as internal consistency (split-half measures), response stability (test-retest), and consistency within and between raters.

   Internal consistency measures are useful if you are gathering information about stake-holders attitudes towards the test or the effect of the test on their attitudes towards the language being tested. Such questionnaires typically include more than one item that is measuring the same issue. One would expect respondents to give comparable responses to items that are measuring the same issue.

   Response stability is also most useful when validating questionnaires. Respondents can be asked to complete the questionnaire on day one and then again the next day. Alderson (1992) used this method in his study of the effect of an exchange programme upon students' language proficiency. He warned, however, that response stability must be checked item by item rather than by aggregating responses across items. Response stability measures can also be used in a modified form for interviews. In this case the respondents would be interviewed twice on consecutive days. The researcher and the respondent could then review the interviews together. Differences in the responses to each question could be discussed in order to establish whether the change in the response had been prompted by a difference in the phrasing of the question or the approach taken by the interviewer. It is important to recall, however, that interviews are a social event and some variability is to be expected and must be borne. The key, nevertheless, lies in minimising the effect of the interpersonal interaction between interviewer and respondent upon the data that is collected.

   Establishing consistency within and between raters is important for the use of checklists and analytical frameworks. It is also important in all aspects of language analysis. To establish intra-rater consistency, judges will need to complete the data collection instrument twice. The stability of the judges' decisions could then be inspected. For instance, if a judge were applying Bachman & Palmer's (1996) Task characteristics framework to a reading test, he/she would need to complete the judgement on two separate occasions (perhaps on consecutive days). His/her judgements could then be compared for consistency. Similarly, if the judge was using a classroom observation instrument he/she would need to complete the checklist twice. In this case, the consistency check would have to be carried out with a video-recorded class. Similar procedures could be applied to establish intra-rater consistency. In this case the completed assessments/observations of two or more different judges would be compared. In both cases, it would be important to interview the judges as well in order to explore inconsistencies in the judgements. It will be important to establish whether any inconsistencies that occur have been caused by changes in the judges' interpretation of what they have been seeing (perhaps a training issue) or by problems with the wording of the instrument.

2. Validity measures such as investigations of content relevance and coverage, and of interpretations of question wording.

   Investigations of content relevance and coverage are useful for questionnaires, checklists, task characteristics frameworks and interviews. For instance, if you were designing a speaking test observation checklist similar to the one designed by O'Sullivan et al. (2002), you could ask expert judges (item writers, teachers etc) to discuss what they would expect the test to include and what they would expect a validation checklist to include. You could then show the judges the actual checklist and ask them to assess the content relevance and coverage of the instrument. This discussion might reveal areas of construct under-representation and/or construct irrelevant items.

Explorations of the way in which respondents interpret questions will help you to establish whether the respondents have understood the question in the way that it was intended. This is particularly useful for questionnaires and interviews but might also be useful in the validation of checklists. In the latter case you want to ensure that your observers have understood the categories they need to gather data under. Clearly, one way of exploring how respondents interpret interview and questionnaire prompts or observation categories would be to conduct a verbal protocol (e.g. Alderson, 1992 and Block, 1998). Alderson (1992) had designed a questionnaire to explore the benefits for their language proficiency of an exchange programme for university students across Europe. Alderson (1992) used verbal protocols to explore respondents' interpretations of the questionnaire items. Block (1998) replicated this methodology in his validation of an end-of-course evaluation form. Block (1998) was particularly interested whether different respondents interpreted the questionnaire items in the same way and also in whether they interpreted the points on the 1-5 rating scale in the same way. Block (1998) reported a high degree of variability in the respondents' interpretations of the questionnaire items and the rating scale. This had implications for Block's ability to aggregate and interpret the questionnaire results.

Foddy (1993:186) suggests an alternative approach to verbal reports, where respondents are asked to rephrase the questions in their own words. You could then analyse these reformulations according to four parameters:

i. fully correct - leaving out no vital parts
ii. generally correct – no more than one part altered or omitted
iii. partially wrong – but indicating that the respondent knew the general subject of the question
iv. completely wrong and no response

(Foddy, 1993: 186)

This approach is interesting because it could be less time-consuming than verbal protocols and might also circumvent some of the problems associated with gathering verbal report data (see 2.1 for this discussion).

Apart from validity, another area of concern for qualitative research is our ability to generalise from the study sample to the wider population. The key to this lies in the representativeness and size of our sample. However, as Lazaraton (1995: 465) argues, even if a result has been established on a large, randomly selected sample, this does not guarantee that it will apply to a particular individual. More importantly, Cronbach (1975) argues that all analyses are context bound:

> Generalizations decay. At one time a conclusion describes the existing situation well, at a later time it accounts for rather little variance, and ultimately is valid only as history.
>
> (Cronbach, 1975: 122)

Cronbach suggests instead that, instead of focusing upon the generalisability of our results, we should make clear the effect of context upon the results, giving "proper weight to local conditions" (1975: 125). He also believes that "any generalization is a working hypothesis, not a conclusion" (1975: 125). These comments are important for they remind us that research is systematic, observant and reflective. It is important to be persuasive and to be seen to have paid attention all the data and to have attempted to account for all of it (rather than just the convenient bits of it). They also highlight the importance of the results having "explanatory power" (Strauss & Corbin, 1998: 267).

## 8. Conclusion

This section of the reference supplement has provided an overview of the range of qualitative methods available for investigating test quality. It has demonstrated the variety of options available and explained the key features of each. In addition, examples of research using the methods have been provided so that you can see how specific qualitative methods have been implemented. The final sub-section (7.1 – 7.6) has also addressed more general issues such as transcription and triangulation of data. The key messages of this section have been:

1. Qualitative methods have enormous power to explain and augment the statistical evidence we might gather to establish test quality.
2. Many of the methods are complimentary and can be used for the triangulation of data sources.
3. It is important to safeguard the validity and generalisability of your data collection methods in order to legitimise the inferences you draw from them.

**References**


Alderson, J.C. (1990) Testing reading comprehension skills (part two): getting students to talk about taking a reading test (a pilot study), Reading in a Foreign Language, 7(1), 465 – 503.

Alderson, J.C. (1992) Validating questionnaires, CRILE Working Papers 15, Lancaster: Department of Linguistics and English Language, Lancaster University.

Alderson, J.C. (2000) Assessing reading, Cambridge: Cambridge University Press.

Alderson, J.C. and Banerjee, J. (2001) Impact and washback research in language testing, in Elder, C., Brown, A., Grove, E., Hill, K., Iwashita, N., Lumley, T., McNamara, T. and O'Loughlin, K. (eds.) Experimenting with uncertainty: essays in honour of Alan Davies, Cambridge: University of Cambridge Local Examinations Syndicate, 150 – 161.

Alderson, J.C. and Hamp-Lyons, L. (1996) TOEFL preparation courses: a study of washback, Language Testing, 13(3), 280 – 297.

Alderson, J.C. and Pižorn, K. (eds.) (2004) Constructing school leaving examinations at a national level – meeting European standards, Ljubljana, Slovenia: The British Council & Državni izpitni center.

Alderson, J.C. and Wall, D. (1992) The Sri Lankan O-Level evaluation project: fourth and final report, Lancaster University.

Allan, A. (1992) Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers, Language Testing, 9, 101 – 122.

Allen, P., Fröhlich, M. and Spada, N. (1984) The Communicative Orientation of Language Teaching: An Observation Scheme, in Handscombe, J., Orem, R.A. and Taylor B.P. (eds) On TESOL '83: The Question of Control, Washington D.C.: TESOL.

Allwright, J. and Banerjee, J. (1997) Investigating the accuracy of admissions criteria: a case study in a British university, CRILE Occasional Report 7, Lancaster: Lancaster University, Department of Linguistics and Modern English Language.

Arnaud, P.J.L. (1984) The lexical richness of L2 written productions and the validity of vocabulary tests, in Culhane, T., Klein-Braley, C. and Stevenson, D.K. (eds.) Practice and problems in language testing, Occasional Papers No. 29, Department of Language and Linguistics, University of Essex, 14 – 28.

Bachman, L.F. (1990) Fundamental considerations in language testing, Oxford: Oxford University Press.

Bachman, L.F., Davidson, F., Ryan, K. & Choi, I.C. (1995) An investigation into the comparability of two tests of English as a foreign language. The Cambridge-TOEFL comparability study, Cambridge: Cambridge University Press.

Bachman, L.F. and Palmer, A.S. (1996) Language testing in practice, Oxford: Oxford University Press.

Banerjee, J.V. (1996) UCLES Report: The design of the classroom observation instruments, unpublished report commissioned by the University of Cambridge Local Examinations Syndicate (UCLES), Cambridge: UCLES.

Banerjee, J.V. (1999) Being an insider – a double-edged sword?, paper presented at the BAAL/ CUP Seminar 1999, Lancaster, U.K.

Banerjee, J.V. (2003) Interpreting and using proficiency test scores, unpublished PhD dissertation, Lancaster University.

Block, D. (1998) Exploring interpretations of questionnaire items, System, 26, 403 – 425.

Borg, W.R., & Gall, M.D. (1983) Educational Research: An Introduction (4th ed.) New York: Longman Inc.

British National Corpus, maintained by the Oxford University Computing Services (http://www.natcorp.ox.ac.uk/)

Brown, A. (1993) The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese, Language Testing, 10(3), 277-303.

Brown, A. (2003) Interviewer variation and the co-construction of speaking proficiency, Language Testing, 20(1), 1 – 25.

Brown, A. and Hill, K. (1998) Interviewer style and candidate performance in the IELTS oral interview, in Woods, S. (ed.) IELTS Research Reports: Volume 1, Sydney, ELICOS, 173 – 191.

Brown, A., and Lumley, T. (1997) Interviewer variability in specific-purpose language performance tests, in Huhta, A.., Kohonen, V., Kurki-Suonio L. and Luoma S. (eds.) Current Developments and Alternatives in Language Assessment, Jyväskylä: Centre for Applied Language Studies, University of Jyväskylä, 137 - 150.

Brown, J.D. and Rodgers, T.S. (2002) Doing second language research, Oxford: Oxford University Press.

Buck, G. (1994) The appropriacy of psychometric measurement models for testing second language listening comprehension, Language Testing, 11(2), 145 – 170

Clapham, C. (1997) IELTS Research Report 3, The British Council, the University of Cambridge Local Examinations Syndicate and the International Development Project for Australian Universities and Colleges, Cambridge.

Clapham, C. (1996) The development of IELTS: A study of the effect of background knowledge on reading comprehension, Cambridge: Cambridge University Press.

Cohen, A (1984) On taking language tests: what the students report, Language Testing, 1(1), 70 – 81

Cohen (1994) English for academic purposes in Brazil: the use of summary tasks, in Hill, C. and Parry, K. (eds.) From testing to assessment: English as an international language, London: Longman, 174 – 204.

Council of Europe (2003) Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF), Strasbourg: Language Policy Division, Council of Europe

Cresswell, J. W. (2003) Research design: qualitative, quantitative, and mixed methods approaches (2nd Edition), Thousand Oaks, CA: Sage Publications.

Cronbach, L. (1975) Beyond the two disciplines of scientific psychology, American Psychologist, 30, 116 – 127.

Denzin, N.K. (1998) The art and politics of interpretation, in Denzin, N.K. and Lincoln, Y.S. (eds)
Strategies of qualitative inquiry, Thousand Oaks, CA: Sage Publications, Inc., 313 - 344.

Dey, I. (1993) Qualitative data analysis: a user-friendly guide for social scientists, London: Routledge.

Dörnyei, Z. (2003) Questionnaires in second language research: construction, administration and processing, Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Eignor, D., Taylor, C., Kirsch, I. and Jamieson, J. (1998) Development of a scale for assessing the level of computer familiarity of TOEFL examinees, TOEFL Research Reports 60, Princeton, NJ: Educational Testing Service.

Feldman, M.S. (1995) Strategies for interpreting qualitative data, Qualitative research methods series 33, Thousand Oaks, CA: Sage Publications, Inc.

Foddy, W. (1993) Constructing questions for interviews and questionnaires: theory and practice in social research, Cambridge: Cambridge University Press.

Fulcher, G. (2003) Testing second language speaking, Cambridge: Polity Press.

Gass, S.M. and Mackey, A. (2000) Stimulated recall methodology in second language research, Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Ginther, A. and Grant, L. (1997) Effects of language proficiency and topic on L2 writing, paper presented at the annual conference for Teachers of English to Speakers of Other Languages, Orlando, Florida, March 1997.

Green, A. (1998) Verbal protocol analysis in language testing research: a handbook, Studies in Language Testing 5, Cambridge: University of Cambridge Local Examinations Syndicate.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B. and Kantor, R. (1996) A study of writing tasks assigned in academic degree programs, TOEFL Research Report No. 54, Princeton, NJ: Educational Testing Service.

Halvari, A. and Tarnanen, M. (1997) Some aspects on using qualitative procedures to ensure comparability across languages within a testing system, in Huhta, A., Kohonen, V., Kurki-Suonio, L. and Luoma, S. (eds.), Jyväskylä: Centre for Applied Language Studies, University of Jyväskylä, 127 – 136.

Hambleton, R. (2001) Setting performance standards on educational assessment and criteria for evaluating the process, in Cizek, G. (ed.) Setting performance standards: concepts, methods and perspectives, Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 89 – 116.
Heritage, J. (1984) Garfinkel and ethnomethodology, Cambridge: Polity.
Herington, R. (1996) Test-taking strategies and second language proficiency: is there a relationship?, unpublished MA dissertation, Lancaster University.
Horák, T. (1996) IELTS impact study project, unpublished MA assignment, Lancaster University.
Hutchby, I. and Wooffitt, R. (1998) Conversation Analysis: An Introduction, Cambridge: Polity Press.
Kane, M.T. (2001) So much remains the same: conceptions and status of validation in setting standards, in Cizek, G.J. (ed.) Setting performance standards: concepts, methods and perspectives, Mahwah, NJ: Erlbaum, 53 – 88.
Kelly, P. (1991) Lexical ignorance: the main obstacle to listening comprehension with advanced foreign language learners, IRAL, 24, 135 – 149.
Kim, S. (2004) A study of development in syntactic complexity by Chinese learners of English and its implications on the CEF scales, unpublished MA dissertation, Lancaster University.
Kirsch, I., Jamieson, J., Taylor, C. and Eignor, D. (1998) Computer familiarity among TOEFL examinees, TOEFL Research Reports 59, Princeton, NJ: Educational Testing Service.
Kormos, J. (1999) Simulating conversations in oral-proficiency assessment: a conversation analysis of role play and non-scripted interviews in language exams, Language Testing, 16(2), 163 – 188.
Laufer, B. (1991) The development of L2 lexis in the expression of the advanced language learner, Modern Language Journal, 75, 440 – 448.
Laufer, B. and Sim, D.D. (1985) Measuring and explaining the reading threshold needed for English for Academic Purposes texts, Foreign Language Annals, 18, 405 – 411.
Lazaraton, A. (1995) Qualitative research in Applied Linguistics: a progress report, TESOL Quarterly, 29(3), 455 – 472.
Lazaraton, A. (2002) A qualitative approach to the validation of oral language tests, Cambridge: UCLES/CUP.
Leech, G., Rayson, P. and Wilson, A. (2001) Word frequencies in written and spoken English: based on the British National Corpus, London: Longman.
Li, W. (1992) What is a test testing? An investigation of the agreement between students' test taking processes and test constructors' presumption, unpublished MA Thesis, Lancaster University.
Low, G. (1996) Intensifiers and hedges in questionnaire rating scales, Evaluation and Research in Education, 2(2), 69 – 79.
Lumley, T. (2002) Assessment criteria in a large-scale writing test: what do they really mean to the raters?, Language Testing, 19(3), 246 – 276.
Marinič, Z. (2004) Test quality, in Alderson, J.C. and Pižorn, K. (eds.) (2004) Constructing school leaving examinations at a national level – meeting European standards, Ljubljana, Slovenia: The British Council & Državni izpitni center, 179 – 192.
Maxwell, J.A. (1992) Understanding and validity in qualitative research, Harvard Educational Review, 62(3), 279 – 300.
Mishler, E.G. (1986) Research interviewing: context and narrative, Cambridge, Mass.: Harvard University Press.
Moser, C.A. and Kalton, K. (1971) Survey Methods in Social Investigation (2nd ed.) Aldershot, Hants.: Gower.
O'Loughlin, K. (1995) Lexical density in candidate output, Language Testing, 12(2), 217-237.
O'Loughlin, K. (2002) The impact of gender in oral proficiency testing, Language Testing, 19(2), 169 – 192.
O'Sullivan, B, Weir, C.J. and Saville, N. (2002) Using observation checklists to validate speaking tasks, Language Testing, 19(1), 33 – 56.
Oppenheim, A.N. (1992) Questionnaire design, interviewing and attitude measurement, London: Pinter Publishers Ltd.

Potter, J. (1996) Discourse analysis and constructionist approaches: theoretical background, in Richardson, J. (ed.) Handbook of qualitative research methods for psychology and the social sciences, Leicester: BPS, 125 – 140.
Potter, J. (1997) Discourse analysis as a way of analysing naturally-occurring talk, in Silverman, D. (ed.) Qualitative research: theory, method and practice, London: Sage Publications Inc., 144 – 160.
Potter, J. and Wetherall, M. (1987) Discourse and social psychology: beyond attitudes and behaviour, London: Sage Publications.
Purves, A.C., Soter, A., Takala, S. and Vähäpassi, A. (1984) Towards a domain-referenced system for classifying assignments, Research in the Teaching of English, 18(4), 385 – 416.
Read, J. (2001) Assessing vocabulary, Cambridge: Cambridge University Press.
Sarig, G. (1987) High level reading tasks in the first and in a foreign language: some comparative process data, in Devine, J., Carrell, P.L. and Eskey, D.E. (eds) Research in reading in English as a second language, Washington, D.C.: TESOL, 105 – 120.
Shohamy, E. (1994) The validity of direct versus semi-direct oral tests, Language Testing, 11(2), 99-123.
Shohamy, E., Donitsa-Schmidt, S. and Ferman, I. (1996) Test impact revisited: washback effect over time, Language Testing, 13(3), 298 – 317.
Silverman, D. (1993) Interpreting qualitative data: methods for analysing talk, text and interaction, London: Sage Publications, Ltd.
Silverman, D. (2001) Interpreting qualitative data: methods for analysing talk, text and interaction, London: Sage Publications Ltd.
Stimson, G.V. (1986) Viewpoint: Place and space in sociological fieldwork, Sociological Review, 34(3), 641 – 656.
Strauss, A. and Corbin, J. (1998) Basics of qualitative research: Techniques and procedures for developing grounded theory, Thousand Oaks, CA: Sage Publications, Inc.
Symon, G. (1998) Qualitative research diaries, in Symon, G and Cassell, C. (eds.) Qualitative methods and analysis in organisational research: a practical guide, London: Sage Publications Inc., 94 – 117.
Taylor, C., Jamieson, J., Eignor, D., and Kirsch, I. (1998) The relationship between computer familiarity and performance on computer-based TOEFL test tasks, TOEFL Research Reports 61, Princeton, NJ: Educational Testing Service.
ten Have (1999) Doing conversation analysis, London: Sage Publications Ltd.
Wall, D. and Alderson, J.C. (1993) Examining washback: the Sri Lankan impact study, Language Testing, 10(1), 41-69.
Weigle, S.C. (1994) Effects of training on raters of ESL compositions, Language Testing, 11(2), 197 – 223.
Weigle, S.C. (2002) Assessing writing, Cambridge: Cambridge University Press.
Wigglesworth, G. (1997) An investigation of planning time and proficiency level on oral test discourse, Language Testing, 14(1), 85 – 106.
Winetroube, S. (1997) The design of the teachers' attitude questionnaires, unpublished report commissioned by the University of Cambridge Local Examinations Syndicate (UCLES), Cambridge: UCLES.
Wolfe-Quintero, K., Inagaki, S. and Kim, H.Y. (1998) Second language development in writing: measures of fluency, accuracy and syntactic complexity, Hawaii: University of Hawaii.
WordSmith Tools, developed by Mike Scott (http://www.oup.com/elt/global/isbn/6890/)