1 Introduction

Rationale

This study explores the equivalence of direct (live) and semi-direct (tape-mediated) speaking tests. This has become an important issue in language testing with the recent advent of semi-direct tests which claim to represent firstly, a valid and reliable substitute for direct procedures in many contexts and secondly, a more standardised and cost-efficient approach to the assessment of oral language proficiency than their direct counterparts. The key question examined in this study is whether or not the two test formats can be considered equivalent in both theoretical and practical terms. This equivalence issue is examined here in the context of the oral interaction component of the ***access:*** test (the Australian Assessment of Communicative English Skills), a 'high stakes' English language test targeted at prospective skilled migrants from non-English speaking backgrounds (NESB).

The ***access:*** oral interaction sub-test was developed in two versions – direct (live) and semi-direct (tape-mediated) – and administered in test centres around the world between 1993 and 1998. The direct version was designed to be used on an individual face-to-face basis (i.e. a single candidate speaking with a trained interlocutor) while the semi-direct version was developed for use in a language laboratory setting where groups of test takers undertake the test simultaneously. Administrators at the overseas test centres were therefore able to make a choice between the two versions based on the human and/or technical resources available to them at any given time. Specifically, this decision depended on first, the number of candidates being tested at each centre; secondly, the technological facilities available (including language laboratories); and thirdly, the availability of suitable interlocutors for the live version.

Since test takers were assigned arbitrarily to either version depending on the location where they undertook the test, it was important that their performance should not be adversely affected by the particular format to which they were allocated. This issue provided the practical motivation for the investigation into the interchangeability of the two versions of the ***access:*** oral interaction sub-test undertaken in this study.

Given the constraints placed on overseas test centres it is important to note at this point that the central validation question did not involve determining

which version of this speaking test was *preferable* but, instead, to what extent the two versions could be considered *equivalent* on the basis of data drawn from test trials. The development of the test is described in more detail in Chapter 2.

Methodological approach

From a theoretical perspective it should be noted that much previous comparability research in language testing has been based on *concurrent validation,* which focuses on the degree of equivalence between test scores. Traditionally, this validation procedure has examined the strength of correlation between scores derived from two tests. High correlations are taken to indicate that the two tests measure the same language abilities while low correlations suggest this is not the case. Many of the empirical studies reported later in this chapter attempt to establish the equivalence of direct and semi-direct speaking tests in this way. However, as Shohamy (1994) convincingly argues, investigating the relationship between test scores provides necessary but insufficient evidence as to whether the same language abilities are being tapped in different tests. She suggests that this issue can only be answered through the more complex process of *construct validation* in which concurrent validation plays an important but nevertheless partial role.

This study therefore attempts to go beyond concurrent validation in order to examine the comparability or equivalence of the direct and semi-direct versions of the ***access:*** oral interaction sub-test. A *case study approach* (Merriam 1988; Yin 1989; Johnson 1992; Nunan 1992) was adopted to carry out the investigation because of first, its holistic focus on the 'bounded system' (i.e. the ***access:*** oral interaction sub-test); secondly, its exploratory, iterative orientation; and thirdly, its capacity to accommodate different philosophical perspectives and research methods (both quantitative and qualitative). This research project was conceived as an *instrumental* case study (Stake 1994) because, in examining the comparability of the live and tape-based versions of this speaking test, it aimed to shed light on the potential equivalence of this and other pairs of direct and semi-direct oral proficiency tests.

In philosophical terms, (as outlined in Chapter 3), an accommodationist stance (Cherryholmes 1992; Lynch 1996) was used to address the research question. This stance enabled the equivalence issue to be investigated from within both the *positivistic* and *naturalistic* research paradigms. Because of its dual emphasis on both product and process and its reliance on both quantitative and qualitative research methods, this strategy eventually allowed for more solidly grounded, valid conclusions than would have been the case if only one paradigm had been used.

The data for the study were collected from two separate trials of this test (December 1992 and June 1994) where candidates undertook both the live and tape-based versions of the ***access:*** oral interaction sub-test.

In the first 'case', the December 1992 trial, the comparability issue was addressed from within a positivistic framework and the focus was on different kinds of *products*, test scores and test taker language output. Firstly, the equivalence of scores obtained by the trial candidates who had completed both versions was examined using multi-faceted Rasch measurement. Secondly, in order to investigate whether the language produced under the two test conditions was comparable, the discourse features of sample audiorecordings from the December 1992 trial were analysed both qualitatively and quantitatively using a framework developed by Shohamy (1994). The focus on test scores and test taker output in this trial yielded important but contradictory evidence in relation to the equivalence issue. This subsequently led to the adoption of another very different perspective from which to address the research question in a subsequent trial.

In the second 'case', the June 1994 trial, the comparability issue was first examined from a naturalistic perspective and the investigation focused on test *processes* including the processes of test design, test taking and rating. This involved tracking the various stages of the trial and gathering a variety of data using observation, interviews and questionnaires. In this case both the data and methods of analysis were mainly qualitative. The test scores from this trial were then analysed quantitatively again using multi-faceted Rasch analyses and the results of selected candidates interpreted using the findings from the previous study of test taking processes. This led to additional quantitative analyses of the test scores from this trial.

By moving back and forward between the positivistic and naturalistic perspectives, therefore, the researcher was able to gather a wide range of evidence to support the conclusions reached in the study. The necessity for this dual perspective will become clearer as the evidence on the validity of the live and tape-based tests unfolds in later chapters.

Structure of the book

The rest of this chapter reviews the literature comparing direct and semi-direct tests of oral language proficiency. After introducing direct, semi-direct and indirect tests of oral proficiency, it discusses the most important theoretical claims made about direct and semi-direct tests and then examines the findings reported in a range of empirical studies comparing the two kinds of tests. Chapter 2 introduces the ***access:*** test in general and the oral interaction sub-test in particular. The comparability of the two versions of the oral interaction sub-test is also briefly examined from the perspective of the relevant test specifications. Chapter 3 describes the methodology used to

empirically investigate the equivalence of the direct and semi-direct versions of the ***access:*** oral interaction sub-test. Chapter 4 examines this issue in relation to the test scores obtained from the first trial held in December 1992 using multi-faceted Rasch measurement. Chapter 5 looks at the comparability question from the perspective of test taker language output on the two versions in the same trial. Chapter 6 explores the test design, test taking and rating processes in a later trial (June 1994) in order to provide a very different perspective on the equivalence of the two versions. Chapter 7 examines the test scores from this second trial again using multi-faceted Rasch analyses. Chapter 8 summarises the findings of the research and then evaluates the usefulness of the various methodologies used in the study to address the main research question and the significance of the findings based on these techniques.

Direct, semi-direct and indirect speaking tests

Clark (1979) provides the basis for distinguishing three distinct types of speaking tests, namely, indirect, semi-direct and direct tests. *Indirect* tests generally refer to those procedures where the test taker is not actually required to speak and belong to the 'precommunicative' era in language testing. Examples of this kind of procedure are the pronunciation tests of Lado (1961) in which the candidate is asked to indicate which of a series of printed words is pronounced differently from others. *Direct* speaking tests, on the other hand, according to Clark (1979: 36) are

> *... procedures in which the examinee is asked to engage in a face-to-face communicative exchange with one or more human interlocutors.*

Direct tests first came into use in the 1950s when the Oral Proficiency Interview (OPI) was developed by the US Foreign Services Institute (FSI). The OPI, as it was originally conceived, is a relatively flexible, unstructured oral interview which is conducted with individual test takers by a trained interviewer who also assesses the candidate using a global band scale. This model has been widely adopted around the world since the 1970s as the most appropriate method for measuring general speaking proficiency in a second language. The Australian Second Language Proficiency Ratings (ASLPR) oral interview developed by Ingram and Wylie (1984) is modelled closely on the original OPI.

In the last decade or so different models of the OPI have evolved. In response to criticisms about the validity and reliability of the original OPI there has been a growing trend towards greater standardisation of the procedure using a range of specified tasks which vary in terms of such characteristics as topic, stimulus, participant roles and functional demands.

An important example of this kind of test is the speaking component of the International English Language Testing System (IELTS), which has been developed by the University of Cambridge Local Examinations Syndicate (UCLES) and is used to assess the readiness of candidates to study or train in the medium of English. The IELTS test can presently be taken in 105 different countries around the world each year. The current speaking sub-test takes the form of a structured interview consisting of five distinct sections which systematically vary the communicative demands made on candidates. These include an introduction where the candidate and interviewer introduce themselves, an extended discourse task in which the candidate speaks at length about a familiar topic, an elicitation task where the candidate is required either to elicit information from the interviewer or to solve a problem, a speculation and attitudes task where the candidate is encouraged to talk about his/her future plans and proposed course of study, and finally a conclusion where the interview is brought to a close (UCLES 1999). UCLES has developed other similar types of speaking tests including the Preliminary English Test, Cambridge First Certificate in English and Certificate of Proficiency in English oral interviews. This more structured, task-based approach to the direct testing of speaking has grown considerably in popularity around the world in recent years. It was also adopted in the development of the direct version of the ***access:*** speaking sub-test (see Chapter 2).

The term *semi-direct* is employed by Clark (1979: 36) to describe those tests which elicit active speech from the test taker

> *... by means of tape recordings, printed test booklets, or other 'non-human' elicitation procedures, rather than through face-to-face conversation with a live interlocutor.*

Normally an audio-recording of the test taker's performance is made and later rated by one or more trained assessors.

Semi-direct tests first appeared during the 1970s and have grown considerably in popularity over the last 25 years, especially in the United States. They represented an early attempt to standardise the assessment of speaking while retaining the communicative basis of the OPI (Shohamy 1994: 101). In addition, they are clearly more cost efficient than direct tests, particularly when administered to groups in a language laboratory, and provide a practical solution in situations where it is not possible to deliver a direct test e.g. where the training and/or deployment of interlocutors is a problem. In recent years they have come under close scrutiny in relation to their validity in particular as we shall see later in this chapter.

Examples of semi-direct procedures used in the US include the Test of Spoken English (TSE) (Clark and Swinton 1979), the Recorded Oral

Proficiency Examination (ROPE) (Lowe and Clifford 1980) and the Simulated Oral Proficiency Interview (SOPI) (Stansfield *et al.* 1990). Examples of semi-direct tests designed in the United Kingdom include the Test in English for Educational Purposes (TEEP) (James 1988) and the Oxford-ARELS Examinations (ARELS Examinations Trust 1989).

Of the three procedures – direct, semi-direct and indirect tests of oral proficiency – indirect tests are generally viewed as the least valid measure of the ability to speak a language precisely because the test taker is not required to speak at all in the course of the test.

Establishing the equivalence of direct and semi-direct tests

This section reviews the most important theoretical arguments and empirical findings to date about the potential equivalence of direct and semi-direct speaking tests in relation to their relative validity, reliability and practicality.

Theoretical claims

Validity

In opening the debate on the equivalence issue Clark (1979) argued that direct tests are the most valid procedures as measures of global speaking proficiency because of the close relationship between the test context and 'real life'. In other words, direct tests more authentically reflect the conditions of the most common form of 'real world' communication, face-to-face interaction. Yet, Clark (1979: 38) also acknowledges that the OPI, the most widely used direct procedure, fails to meet these conditions in two important respects. First, there is the problem of the interviewer:

> *In the interview situation, the examinee is certainly aware that he or she is talking to a language assessor and not to a waiter, taxi driver, or personal friend.*

Secondly, the language elicited in an interview is unlikely to reflect the discourse of 'real-life' conversation. In particular, the fact that the interviewer controls the interview means that the candidate is normally not required to ask questions.

Hughes (1989) and van Lier (1989) also challenge the validity of the oral interview in terms of this asymmetry which exists between the interviewer and the candidate. Hughes (1989: 104) points out that in an oral interview 'the candidate speaks as to a superior and is unwilling to take the initiative'. Consequently, only one style of speech is elicited, and certain functions (such as asking for information) are not represented in the candidate's performance.

Hughes recommends the inclusion of tasks such as role plays and discussions as ways of varying the type of interaction, although the underlying asymmetry between interviewer and candidate may not be automatically removed by simply incorporating other tasks in which the participants seem more equal.

Van Lier pursues a stronger version of this argument. He questions whether an interview can validly serve the purpose of assessing oral proficiency by contrasting the essential features of conversations and interviews. An interview, in van Lier's (1989: 496) terms, is distinguished by 'asymmetrical contingency':

> *The interviewer has a plan and conducts and controls the interview largely according to that plan.*

On the other hand, a conversation, van Lier (1989: 495) contends, is characterised by

> *face-to-face interaction, unplannedness (locally assembled), unpredictability of sequence and outcome, potentially equal distribution of rights and duties in talk, and manifestation of features of reactive and mutual contingency.*

The emphasis in an interview is on the successful elicitation of language rather than on successful conversation. Van Lier (1989: 505) calls for research into whether or not conversation is the most appropriate vehicle to test oral proficiency. If so, he argues,

> *we must learn to understand the OPI, find out how to allow a truly conversational expression of oral proficiency to take place, and reassess our entire ideology and practice regarding the design of rating scales and procedures.*

If direct tests, particularly oral interviews, can be criticised for their lack of authenticity then, at face value, semi-direct tests are even more open to this charge. Clark (1979: 38), for instance, argues that they

> *require the examinee to carry out considerably less realistic speaking tasks (than direct tests) – such as responding to tape-recorded questions, imitating a voice model, or describing pictures aloud – which, although they do involve active speaking, represent rather artificial language use – situations which the examinee is not likely to encounter in a real-life (i.e. non-test) setting.*

However, it should be noted that such 'artificial' tasks as 'describing pictures aloud' have also been used in some direct tests including the live version of ***access:*** oral interaction sub-test (see Chapter 2).

Underhill (1987: 35) is also strongly critical of the lack of authenticity in semi-direct tests:

> *There are few situations in the real world in which what the learner says has absolutely no effect on what he hears next.*

Secondly, he suggests, there is the problem that the assessor misses visual aspects of the candidate's communication in a semi-direct test since their judgement is normally based on an audio-recording of the test performance. Thirdly, while a direct test can be lengthened or directed more carefully if the interviewer considers the speech sample produced by the candidate to be inadequate for assessment purposes, this is not the case in a semi-direct test where the amount of response time allowed is 'set' in advance. Lastly, speaking to a microphone rather than another person may be unduly stressful for some candidates, especially if they are unused to a language laboratory setting. Possible means of reducing their anxiety include giving instructions in the native language, or in written form, or by ensuring that all test takers are familiar with the system in advance.

Both Clark (1979) and Underhill (1987) therefore clearly favour the use of direct tests over their semi-direct counterparts, at least for measuring general speaking proficiency. Clark (1979: 38) contends that

> *the face-to-face interview appears to possess the greatest degree of validity as a measure of global speaking proficiency and is clearly superior in this regard to both the indirect (non-speaking) and semi-direct approaches.*

Clark (1979: 39) suggests that semi-direct tests lend themselves better to what he calls 'diagnostic achievement tests' which measure discrete aspects of speaking performance such as vocabulary items and syntactic patterns, (although this seems a rather reductive view of the potential use of this kind of test). In general, he argues against using either test type for 'cross purposes', i.e. for either obtaining detailed achievement information using a direct test or measuring global proficiency using a semi-direct test. However, Clark (1979: 48) also suggests that:

> *semi-direct tests may be proposed as second-order substitutes for direct techniques when general proficiency measurement is at issue but it is not operationally possible to administer a direct test. In these*

> *instances, it is considered highly important to determine – through appropriate experimental means – a high level of correlation between the two types of instruments when used with representative examinee groups.*

In accordance with the traditional requirements for concurrent validation (Alderson *et al*., 1995: 178) a correlation of 0.9 or higher is argued to be the appropriate level of agreement at which test users could consider 'the semi-direct testing results closely indicative of probable examinee performance on the more direct measures' (Clark 1979: 40). However, a high correlation between scores obtained from direct and semi-direct tests of oral proficiency does not in itself constitute sufficient evidence that a semi-direct test can be validly substituted for a direct one: the two kinds of tests may not be measuring the same construct. In other words, they could be assessing different components of the oral proficiency trait. The inadequacy of concurrent validation is a central issue in this study and its limitations will be examined more closely later in this chapter in relation to empirical studies previously carried out on the equivalence of direct and semi-direct tests.

Finally, while Clark's (1979) suggestion that direct tests are preferable because they generally approximate 'real-life' communication more closely than semi-direct tests is reasonable (albeit perhaps rather simplistic – see the discussion of the study by Hoejke and Linnell (1994) later in this chapter), he fails to articulate precisely which speaking skills are tapped in the two test formats. In a later publication Clark (1986: 2) is more explicit about what is lacking in semi-direct tests:

> *interactive discourse-management aspects of the student's overall speaking proficiency cannot be readily elicited (or by the same token, effectively assured) through semi-direct techniques.*

This limitation notwithstanding, Clark (1986: 2) is now more optimistic that the semi-direct format

> *can serve to validly and efficiently measure many of the other performance aspects that constitute overall speaking proficiency.*

He argues that this is particularly true of 'proficiency-oriented semi-direct tests' which attempt to approximate as closely as possible the '... linguistic content and manner of operation' as well as the scoring procedures of a live interview.

Van Lier (1989: 493) adopts a less equivocal position than Clark (1986). He considers face-to-face direct tests to be, in principle, more valid than other test formats including semi-direct tests in most circumstances since

> *face-to-face talk is to be regarded as the unmarked form of interaction, and communicating by telephone or speaking into a microphone as marked forms of interaction.*

He argues that proficiency in these marked forms of communication is an advanced skill which should only be tested in special instances:

> *Hence, although remote interaction may be part of performance testing for specific groups of learners, it would appear to be an unfair, that is invalid, measure of general oral proficiency.*

While 'remote communication' may be more difficult for some test takers, this may not necessarily be the case for other people unaccustomed to face-to-face interaction. However, if different speaking abilities do underlie these two kinds of communication then the interchangeability of direct and semi-direct tests of oral proficiency is left in doubt.

Reliability

While semi-direct tests have been typically viewed as inferior to direct tests in relation to validity they are often seen as possessing potentially stronger reliability. Hughes (1989) argues that the chief advantages of semi-direct procedures are the uniformity of their elicitation procedures and the increased reliability which is likely to flow from such standardisation. This uniformity is inevitably placed under threat in direct tests because of interviewer variability. As Lazaraton (1996: 154) suggests,

> *[t]he potential for uneven interviewer performance in a face-to-face interview is one reason that [semi-direct tests] are so appealing i.e. they remove the variability that a live interlocutor introduces.*

This is particularly true of the relatively unstandardised OPI where the content and form of the questions posed to the test taker can vary considerably from one interview to another.

This lack of standardisation can then have adverse effects on test performance and reliability of scoring. Underhill (1987: 31), for example, points out that, in an oral interview, the lack of script or set tasks gives this procedure its flexibility and yet

> *this flexibility means that there will be a considerable divergence between what different learners say, which makes a test more difficult to assess with consistency and reliability.*