The equivalence of direct and semi-direct speaking tests

Kieran J. O’Loughlin
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Series Editor’s note</td>
<td>vii</td>
</tr>
<tr>
<td>Preface</td>
<td>ix</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>xi</td>
</tr>
<tr>
<td><strong>Chapter One</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>Rationale</td>
<td>1</td>
</tr>
<tr>
<td>Methodological approach</td>
<td>2</td>
</tr>
<tr>
<td>Structure of the book</td>
<td>3</td>
</tr>
<tr>
<td>Direct, semi-direct and indirect speaking tests</td>
<td>4</td>
</tr>
<tr>
<td>Establishing the equivalence of direct and semi-direct tests</td>
<td>6</td>
</tr>
<tr>
<td>Research findings</td>
<td>12</td>
</tr>
<tr>
<td>Conclusion</td>
<td>31</td>
</tr>
<tr>
<td><strong>Chapter Two</strong></td>
<td></td>
</tr>
<tr>
<td><strong>The access: test</strong></td>
<td>33</td>
</tr>
<tr>
<td>Background to the test</td>
<td>33</td>
</tr>
<tr>
<td>Development of the test</td>
<td>34</td>
</tr>
<tr>
<td>The oral interaction sub-test</td>
<td>35</td>
</tr>
<tr>
<td>Conclusion</td>
<td>45</td>
</tr>
<tr>
<td><strong>Chapter Three</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Methodology</strong></td>
<td>47</td>
</tr>
<tr>
<td>The instrumental case study approach</td>
<td>47</td>
</tr>
<tr>
<td>Research design of the study</td>
<td>55</td>
</tr>
<tr>
<td>Summary</td>
<td>61</td>
</tr>
<tr>
<td><strong>Chapter Four</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Trial 1: The analysis of test scores</strong></td>
<td>62</td>
</tr>
<tr>
<td>Overview</td>
<td>62</td>
</tr>
<tr>
<td>The December 1992 trial</td>
<td>62</td>
</tr>
<tr>
<td>A preliminary study</td>
<td>65</td>
</tr>
<tr>
<td>The current study</td>
<td>69</td>
</tr>
<tr>
<td>Conclusion</td>
<td>79</td>
</tr>
</tbody>
</table>
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 audiotapes 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 forth 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
1 Introduction

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
1 Introduction

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,

[the 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.
Underhill also contends that the more predictable content of candidate output in semi-directs tests means that the scoring criteria can be more easily and accurately constructed. This, he claims, is likely to yield more reliable results. In direct tests where the interviewer also acts as the rater (such as the OPI) there is also some doubt cast over whose performance is actually being assessed. James (1988: 116) suggests that the situation

... can develop into a catechesis often with examiners’ marks reflecting their satisfaction with their own performances rather than those of the candidates.

In general, therefore, semi-direct tests are believed to provide a more reliable measure of oral proficiency than their direct counterparts.

Practicality

It is frequently claimed that semi-direct tests offer a more practical alternative to the assessment of speaking proficiency than direct tests. Underhill (1987) lists several important advantages which semi-direct tests possess in this regard. First, since groups of candidates can be tested simultaneously in language laboratories using this format, a semi-direct test can be conducted more economically and efficiently. Secondly, the marking does not have to be done in real time since the candidate’s performance is audio-recorded and can take place whenever and wherever it is convenient to do so (although it should be noted that the same provision could be built into a direct test). Thirdly, its fixed structure allows assessors to listen to the tapes more quickly by fast-forwarding past instructions and longer task stimuli. Fourthly, tape-mediated tests can be useful where there are logistical problems in obtaining suitable interviewers. Finally, it is undeniable that most forms of direct procedures are more expensive than their semi-direct counterparts because of the higher costs incurred in the selection and training of interviewers and assessors and the administration of the test.

On the other hand, Underhill notes that technical problems can result in poor quality recordings or even no recording at all when a semi-direct format is adopted. Of course, either of these eventualities can occur when direct tests are being recorded for later assessment by raters other than the interviewer. However, live tests where the interviewer carries out the assessment obviously avoid this problem.

At this point it is important to note that in much of the preceding discussion the term ‘direct test’ is defined solely in relation to the OPI no doubt because of its dominance in the world of proficiency testing until very recently. The prototypical OPI is very different from most semi-direct procedures not simply in terms of whether the candidates speak to a microphone or another
1 Introduction

person but also in terms of its structure and its degree of standardisation. Taken together these differences make it extremely difficult to compare the two kinds of tests.

Research findings

This section reviews the empirical research literature on the comparability of direct and semi-direct tests of oral proficiency.

Concurrent validation

Following Clark (1979), attempts to establish the equivalence of these two types of speaking test have relied primarily on concurrent validation, at least until very recently. This form of validation involves correlating the test scores obtained from the same group of test takers, who have undertaken two different tests. If the correlation between the scores is strong then it has been traditionally inferred (often prematurely) that the two tests are measuring the same ability. Concurrent validation also assumes that one of the tests is a valid measure of the language ability in question and can therefore serve as the ‘criterion behaviour’ (Bachman 1990: 249). In most of the research reported to date the direct test – normally the OPI – fulfils this function. It is assumed that it is the semi-direct test whose validity needs to be established in this process.

The OPI and the scale which is used to rate it have been in use in various forms since the 1950s. The OPI was originally developed by the US Foreign Service Institute which is responsible for the training of diplomats and foreign service officers in foreign languages. It was adapted by the American Council for the Teaching of Foreign Languages (ACTFL), the Educational Testing Service (ETS) and the Interagency Language Roundtable (ILR) in the 1970s and became known therefore as the main assessment tool of the ‘AEI’ proficiency movement (Lowe 1988). The OPI, and the scale on which it is scored, have been widely accepted as a standard for assessing oral proficiency in a foreign language in the US and other countries (Bachman 1991). It is also the precursor of the Australian Second Language Proficiency Ratings (ASLPR) (Ingram and Wylie 1984).

The prototypical OPI consists of a face-to-face interview conducted in the target language by a trained interlocutor (who usually also carries out the assessment) and can include a role play segment. The topics, language input and timing of the interview are adjusted according to the candidate’s perceived proficiency which is probed using questions requiring increasingly more complex responses. The topics are chosen by the interlocutor from a range of possible options as specified in the test manual. The OPI usually ends with a ‘wind down’ phase consisting of one or more easy questions designed
to put the examinee at ease and to facilitate the ending of the test. The candidate’s performance is scored holistically using either the ILR or ACTFL/ETS scales. The ILR scale consists of 11 levels of general oral proficiency ranging from ‘0’ (no ability to communicate effectively in the language) to ‘5’ (functioning as an educated native speaker). It includes five mid-points (i.e. 0+, 1+, 2+, 3+, 4+) for performances that surpass the requirements for a given level but which fail to reach the next level. However, these mid-points are not defined. The ACTFL/ETS scale, which is derived from the ILR scale, describes different proficiency levels beginning at ‘Novice’, which consists of three sub-levels, moving up to ‘Intermediate’, also comprising three sub-levels, ‘Advanced’, with two sub-levels and finally ‘Superior’. Tables 1.1 and 1.2 provide the headings from the official descriptions of proficiency at the various levels on the ILR and ACTFL/ETS scales respectively. Table 1.3 shows the relationship between the ILR scale and the ACTFL/ETS scale (Lowe 1987; Clark and Clifford 1988).

Table 1.1
The ILR scale

<table>
<thead>
<tr>
<th>LEVEL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEVEL 1</td>
<td>Elementary proficiency: able to satisfy routine travel needs and minimum courtesy requirements.</td>
</tr>
<tr>
<td>LEVEL 2</td>
<td>Limited working proficiency: able to satisfy routine social demands and limited work requirements.</td>
</tr>
<tr>
<td>LEVEL 3</td>
<td>Minimum professional proficiency: able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social and professional topics.</td>
</tr>
<tr>
<td>LEVEL 4</td>
<td>Full professional proficiency: able to use the language fluently and accurately at all levels normally pertinent to professional needs.</td>
</tr>
<tr>
<td>LEVEL 5</td>
<td>Native or bilingual proficiency: speaking proficiency equivalent to that of an educated native speaker.</td>
</tr>
</tbody>
</table>
Introduction

Table 1.2
The ACTFL/ETS scale

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOVICE</td>
<td>The novice level is characterised by an ability to communicate minimally with learned material.</td>
</tr>
<tr>
<td>INTERMEDIATE</td>
<td>The intermediate level is characterised by an ability to create with the language by combining and recombining learned elements, though primarily in a reactive mode; initiate, minimally sustain, and close in a simple way basic communicative tasks; and ask and answer questions.</td>
</tr>
<tr>
<td>ADVANCED</td>
<td>The advanced level is characterised by an ability to converse in a clearly participatory fashion; initiate, sustain, and bring to closure a wide variety of communicative tasks, including those that require an increased ability to convey meaning with diverse language strategies due to a complication or an unforeseen turn of events; satisfy the requirements of school and work situations; and narrate and describe with paragraph-length connected discourse.</td>
</tr>
<tr>
<td>SUPERIOR</td>
<td>The superior level is characterised by the ability to participate effectively in most formal and informal conversations on practical, social, professional, and abstract topics; and support opinions and hypothesise using native-like discourse strategies.</td>
</tr>
</tbody>
</table>

Table 1.3
Relationship between the ILR scale and the ACTFL/ETS scale

<table>
<thead>
<tr>
<th>ILR</th>
<th>ACTFL/ETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 and above</td>
<td>Superior</td>
</tr>
<tr>
<td>2+</td>
<td>Advanced Plus</td>
</tr>
<tr>
<td>2</td>
<td>Advanced</td>
</tr>
<tr>
<td>1+</td>
<td>Intermediate-High</td>
</tr>
<tr>
<td>1</td>
<td>Intermediate-Mid &amp; Low</td>
</tr>
<tr>
<td>0+</td>
<td>Novice-High, Mid &amp; Low</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The OPI and its immediate successors have dominated the landscape of speaking tests for the last 25 years. And yet, periodically, there have been attempts to find a more standardised, cost-effective but still sufficiently valid and reliable assessment instrument to replace it. Semi-direct tests appeared to many language testing experts to have the potential to provide such an alternative.
The earliest attempt to devise a semi-direct test in the US was the TSE (Clark and Swinton 1979). The TSE was introduced by the Educational Testing Service (ETS) in 1980 to meet a general need for an international standardised test of oral English proficiency. It is now used in over 80 countries around the world, often as an adjunct to the Test of English as a Foreign Language (TOEFL), which does not have a speaking component. The TSE has also been widely used by universities in the US to assess the oral proficiency of NESB international teaching assistants (ITAs). The test is tape-based, conducted entirely in English and of approximately 20 minutes’ duration. It consists of six discrete tasks including answering simple personal questions, reading aloud, sentence completion, narration using a sequence of pictures, description and discussion of a topical issue. The test is scored for pronunciation, grammar, fluency and an overall category, comprehensibility, which represents a more general assessment of the test taker’s oral proficiency.

In an early study Clark and Swinton (1980: 18) found that scores on the TSE and OPI obtained by a group of foreign teaching assistants (N = 134) correlated at $r = 0.80$. This result suggested to them that the TSE was ‘a reasonable alternative to the FSI interview when it is not possible to carry out face-to-face testing’ even though it is lower than the figure of 0.9 originally proposed by Clark (1979) to be the acceptable level of agreement at which a semi-direct test might be substituted for a direct test.

In a more recent study Southard and Sheorey (1992: 54) attempted to establish whether a rated interview ‘could serve as a substitute for a standardised measure such as the TSE’. This study is unusual insofar as it uses a semi-direct measure of oral proficiency as the ‘criterion behaviour’ (Bachman 1990: 249) against which the concurrent validity of the direct interview procedure is investigated.

In Southard and Sheorey’s study the TSE results of 19 ITAs in the department of English at Oklahoma State University were compared with their performance on a direct interview test. The interview consisted of a structured conversation between individual ITAs and a panel of five judges including two experienced ESL professionals and three ‘naïve’ assessors (the director of the English department’s freshman composition program and two undergraduate students). Candidates were then rated by each member of the panel on a five point scale in the following categories: pronunciation, grammar, vocabulary, auditory comprehension and overall communicative competence.

On the basis of statistically significant (in most cases) but still relatively low correlations between scores on the various criteria used in the two tests (ranging from 0.19 to 0.89) Southard and Sheorey (1992: 62) suggest that
a well-designed, on-campus interview conducted by experienced ESL teachers who are given structured training in evaluating oral proficiency can be used as an alternative instrument if a standardised test like the TSE is not readily available.

However, given the mainly modest correlation figures reported and the limitations of using such evidence to infer that the two tests are measuring the same abilities, this conclusion appears rather dubious. It is also surprising to note here, unlike in most other recent studies, that a semi-direct procedure is assumed to be valid from the outset and that at no stage of this study is this assumption about the TSE seriously called into question. In the literature generally, it is normally the validity of direct measures of oral proficiency which is taken for granted and that of semi-direct alternatives which needs to be established.

Another early semi-direct test developed in the US, this time designed to assess foreign language oral proficiency, was the ROPE (Lowe and Clifford 1980). The ROPE test more closely resembled the OPI than did the TSE in several important ways. In the ROPE test candidates listen to a series of tape-recorded questions in the target language and provide responses ranging from ‘yes/no’ to detailed expressions of opinion. Unlike the TSE, there are no written or visual stimuli in this test. The ROPE is scored on the ILR scale from 0+ to 5 (see Table 1.1 above). While the ROPE more closely approximated to the OPI in terms of the kinds of questions used and the method of scoring than the TSE, an important limitation was that not all candidates could understand the questions since they were presented in the target language (Stansfield et al. 1990). Lowe and Clifford report a correlation of $r = 0.90$ between scores on the OPI and ROPE across several languages including French, German and Spanish, which meets the level of agreement stipulated by Clark (1979) for test substitution. However, this result should be viewed cautiously given that the total number of subjects was only 27. More importantly, as suggested above, this result alone may not be sufficient to conclude that the two kinds of tests are equivalent.

These studies therefore clearly illustrate the exclusive reliance which has been placed on concurrent validation to establish the equivalence of direct and semi-direct assessment procedures until very recently. As we shall see, this approach was also favoured, at least initially, in comparing the OPI with a more recently developed semi-direct foreign language test, the Simulated Oral Proficiency Interview test (SOPI).

The impetus for the development of the SOPI in the late 1980s was created by problems relating to the testing of less commonly taught languages in the US such as Chinese and Portuguese. In these instances the prohibitive costs and practical difficulties involved in training and deploying suitable
interviewers for the OPI meant that a different test format needed to be created, one that could be administered on a one-off basis to individual candidates as well as to large groups throughout the country where required. A semi-direct test seemed to have the potential to meet these needs.

The first SOPI was developed by Clark and Li (1986) (although not labelled as such) at the Center for Applied Linguistics (CAL), Washington, DC to assess oral proficiency in Chinese. Stansfield and Kenyon (1988) later used this model (formally naming it the simulated oral proficiency interview or SOPI) to create a similar test in Portuguese. This was followed by the development of tests in other less commonly taught languages including Hebrew (Shohamy et al. 1989), Hausa and Indonesian (Stansfield and Kenyon 1989). All of these tests were developed at CAL.

Interestingly, the development of the SOPI also reflected in its design increasing concerns emanating from language testing research about the validity and reliability of the OPI. In terms of its validity Shohamy (1983), for instance, queried whether a single type of interaction (i.e. an interview) was sufficient to assess oral proficiency. Perrett (1987) argued that while the oral interview may enable assessment of the candidate’s phonological and lexico-grammatical skills and some aspects of discourse competence, it does not provide sufficient information about the candidate’s control over an adequate range of text types, speech functions and exchange structures. Raffaldani (1988) suggested that the interview format is the main reason why the OPI fails to elicit some important aspects of communication: a limited number of speech functions are sampled and so candidates have little opportunity to display either discourse or sociolinguistic competence. As noted earlier in this chapter, van Lier (1989) also questioned whether an interview was the most suitable vehicle for the measurement of oral proficiency since it largely fails to tap the test taker’s conversational ability. In terms of reliability, Shohamy (1983) suggested that OPI scores were affected by a range of contextual variables including topic, type of interaction and interviewer behaviour. Each interview is therefore different because these variables are not tightly controlled.

There was growing interest, therefore, amongst language testing experts in exploring the potential of semi-direct tests to enhance the reliability and validity of speaking tests by controlling such variables and including a range of task types (Shohamy 1994). Thus, by the mid-1980s, they were no longer necessarily conceived as merely ‘second order substitutes for direct techniques’ (Clark 1979: 48) but as potentially more valid and reliable than their direct counterparts.

The SOPI differed from the OPI in several important respects. As in the OPI, there is an initial ‘warm up’ phase where the candidate is asked a number of simple personal background questions. The rest of the test consists of a series of set tasks (in contrast to the more open-ended question and answer...
structure of the OPI) which elicit oral discourse through the use of both aural and visual stimuli. These tasks include giving directions using a map, describing a picture, narrating a story based on a picture sequence, talking about topical subjects and lastly, responding to situations in which the communicative tasks and the audience are specified. In addition, all tasks on the SOPI (with the exception of the ‘warm up’ phase) are normally read aloud in English on tape and written in English in a test booklet as well as in the target language. As in the OPI, the test becomes progressively more demanding as the test continues: each task is intended to probe or verify a higher level of proficiency until the final ‘wind down’ stage where candidates are asked one or more simple questions. Unlike the OPI, the SOPI is assessed retrospectively by trained raters using the audio-taped recording of the candidate’s test performance. Like the OPI, however, the SOPI is assessed holistically using the ACTFL/ILR scale, which is grounded in a view of language proficiency as a unitary ability. In other words, the components associated with the various points on the rating scale are considered together rather than individually (Stansfield 1991; Stansfield and Kenyon 1992a; Stansfield and Kenyon 1992b).

On the basis of research carried out at CAL in Washington, DC, Stansfield (1991) suggested that the SOPI had shown itself to be a valid and reliable substitute for the OPI. In relation to a comparison of scores on the two kinds of tests, Stansfield reported high Pearson correlations (\(r\)) in the range of 0.89–0.95 on the OPI and SOPI in Chinese, Portuguese, Hebrew and Indonesian. On the basis of these findings Stansfield (1991: 206) argued that

\[ \text{... the SOPI correlates so highly with the OPI that it seems safe to say that both measures test the same abilities.} \]

However, it should be noted that the numbers of test takers under investigation in each case were again relatively small, ranging from 10 to 30, and the correlation figures should therefore be regarded with some caution.

A more fundamental problem with Stanfield’s conclusion is that, as previously suggested, high correlations between scores on the two kinds of tests provide necessary but insufficient evidence that they measure the same ability: it would also be desirable, for instance, to show that the OPI and SOPI scores in these studies were not consistently related to measures of other language or non-language abilities in order to provide a sounder basis for this conclusion (Bachman 1990: 250). Where high correlations exist between the two tests, it may be possible to safely predict OPI scores from SOPI scores but this does not necessarily imply that the same ability is being measured. Another difficulty with concurrent validation as noted previously is that it assumes that one of the tests is a valid measure of the language ability in
question and can therefore serve as the ‘criterion behavior’. However, as Bachman (1990: 249) suggests:

\[
\text{without independent evidence supporting the interpretation of the criterion as an indicator of the ability in question, there is no basis for interpreting a correlation with that criterion as evidence of validity.}
\]

Given the widespread controversy about the validity of the OPI (see above) it would appear unwise to draw any conclusion about the validity of the SOPI simply on the basis of high correlations between scores on the two tests.

At the statistical level, another limitation of most correlation co-efficients (including the Pearson correlation used in Stansfield’s (1991) study), is that they provide a measure of linearity rather than equivalence between two sets of scores. It is still possible, therefore, that candidates may have performed systematically better on either the OPI or the SOPI despite the strength of the relationship between the scores. The high correlations, therefore, may allow the score of a candidate on either test to be fairly confidently predicted from the other but do not necessarily indicate that the two tests were equally difficult. The inclusion of descriptive statistics in Stansfield’s study, particularly the means, variances and standard deviations of the OPI and SOPI scores for each language, would have assisted in providing this information.

There is an additional problem with the interpretation of high correlations between the scores on the two tests. Unless the correlation is perfect in the positive sense (i.e. +1), there will be candidates for whom the level of performance differed from one test to the other. A correlation of 0.9 indicates that this is the case for approximately 20% of test takers. Thus, in Stansfield’s study, it appears that the results for a significant minority of candidates in each of the four languages could not be predicted from one test to the other. It may be that the performance of these people was affected by the two different test formats (i.e. live or tape-mediated), although there may have been other factors influencing them as well, such as a positive or negative practice effect resulting from taking the two tests within a short space of time.

In a later study, Stansfield and Kenyon (1992b) used generalisability theory (also known as G-theory) to further explore the issue of score comparability using the same data as in Stansfield’s (1991) study. Generally low levels of subject by test interaction were found for candidates who had undertaken the two kinds of tests in all four languages. However, the results of this analysis did suggest a tendency for some candidates to perform differentially on the two test formats in three of the five studies undertaken, a finding which supports the criticism of the correlational results in the earlier study (see above). Unfortunately, Stansfield and Kenyon (1992b: 356) were unable to confirm from their analysis using G-theory whether this indicated
that ‘many examinees deviated a little in their performance on the two tests or whether a few examinees deviated a lot, or some combination in-between’.

Stansfield also examines the comparative characteristics of the OPI and SOPI tests in the light of his empirical findings based on test scores. He argues that the SOPI may be more reliable than the OPI first because ‘the OPI requires that each examinee be given a unique interview, whereas the format and questions on a SOPI are invariant’ (Stansfield 1991: 202). As a result, raters reported that it was often easier to arrive at a decision on a score in the case of the SOPI. Secondly, the greater length of speech sample produced in the SOPI (typically 20–23 minutes versus 15 minutes on the OPI) may make for more accurate judgements about a candidate’s proficiency level. Thirdly, the fact that the SOPI is recorded means that it can be assessed by the most reliable rater while the OPI is sometimes rated by interviewers who may not always be the most accurate judges.

In relation to the issue of validity, Stansfield (1991: 203) suggests that one important problem with the OPI is that the candidate’s performance may be significantly affected by the skill of the interviewer whereas the SOPI offers the same quality of language input to each candidate. This has since been identified as an important source of measurement error in speaking tests.

In addition, Stansfield (1991: 204) argues that the reason the OPI and SOPI correlate so highly may be because ‘neither format produces a “natural” or “real-life” conversation’. Even in the OPI, he contends, both interviewer and candidate understand that ‘it is the examinee’s responsibility to perform. Little true interaction takes place’ (Stansfield 1991: 205). However, such a conclusion needs to be empirically investigated by examining data other than test scores (such as the discourse produced under the two test conditions) to discover whether, in fact, this is indeed the case.

Notwithstanding his concerns about the reliability and validity of the OPI, Stansfield (1991: 207) still views it as ‘potentially the more valid and reliable measure when carefully administered by a skilled interviewer and rated by an accurate rater’. On the other hand, given that both of these conditions are not always met in the case of the OPI, Stansfield considers that the OPI may be more suitable for placement and program evaluation purposes and the SOPI more appropriate when important decisions are to be based on test scores given the high degree of ‘quality control’ it offers. However, this line of reasoning is suspect since it may not always be possible to gauge in advance the importance of the results from any assessment procedure for the life chances of test takers in the short or long term.

From a different perspective, Stansfield appears not to have considered possible sources of variability arising from the semi-direct test format. For instance, candidates may vary considerably in their experience with language laboratories, particularly in a test context. Some test takers may also require