Notes on Assessing Speaking

Barry O'Sullivan

1. Introduction

With all language skill testing, it is necessary to clearly state the specifications before beginning to write the test (just as the previous sections have stated). In his very pragmatic 'Understanding and Developing Language Tests' Cyril Weir (1993) presents an approach to skills testing which calls for the identification of a series of language operations which the skill in question will entail. In addition to this list, Weir suggests that we need a set of conditions under which the task will be performed. Together these will allow the test writer to establish the parameters of the task or tasks to be employed in the test. Weir (2005) updated his original ideas and this later framework adds greatly to our conceptualisation of how all four skills can be tested and validated. In the following sections we will be looking at the testing of the four language skills in terms of Weir's 2005 framework.

It is commonly believed that tests of spoken language ability are the most difficult. While I will not attempt to argue against that view here, it is important to recognise the great improvements in the area that have been made over the last few decades. There remain, of course, a number of areas of great concern to the test writer, most notably construct definition, predictability of task response (task description), interlocutor effect, the effect of characteristics of the test-taker on performance, rating-scale validity and reliability, and rater reliability. McNamara's (1996) model of the relationship of proficiency to performance (see page 17) clearly highlights all of these 'trouble spots'. He is essentially saying that until we can be certain that we fully understand all the elements contained in his model we can not claim to be in a position to create a totally valid or reliable performance test. I do not believe that McNamara really expects that all elements can be 'fully' understood (clearly they cannot), and so — as in all other areas of testing (language or other) — we can only attempt to do our utmost to create tests which reflect what we *do* know, and to limit the inferences we draw from scores awarded on these tests on this basis.

There has been relatively little, though very valuable research done in recent years into those aspects of spoken language testing referred to above.

- Berry (2004), Kunnan (1995) and Purpura (1999) have explored characteristics of the test-taker
- Foster and Skehan (1996; Foster and Skehan, 1999) have looked at the task, as have Norris et al. (1998)
- O'Sullivan & Porter (1996), O'Sullivan (O'Sullivan, 1995; 2000; O'Sullivan, 2002; O'Sullivan, 2006), Brown (2003) and O'Loughlin (2002) have investigated the interlocutor effect;
- North (1995) and Fulcher (1996) have focused on rating scales;
- Wigglesworth (1993), McNamara (1996), McNamara & Lumley (1997) and O'Sullivan (2000) have examined rater performance.

One aspect of speaking that has received some interest in the past decade is the issue of planning ((Foster and Skehan, 1996; Foster and Skehan, 1999; O'Sullivan et al., 2004; Ortega, 1999; Wigglesworth, 1997). It appears clear at this point in time that the provision of planning time is one of the key parameters that affects task difficulty in speaking – though according to the findings of O'Sullivan et al (2004) these effects may well be different for students at different levels of overall ability.

While the difficulties described above all relate to theoretical issues, there are also a great number of practical considerations, which make this type of test more difficult to administer than others. The first of the considerations is the sheer complexity of the logistics involved. A test recently administered in Turkey, required that 400 candidates be tested in pairs in a single day. This meant that there were 200 individual tests, each lasting 15 minutes, with a 5 minute turn-around time. This translates into 67 hours of testing, and with each test requiring two administrators (and interviewer and an observer) who could not be expected to work for more than 6 hours, the total number of rooms required was 12 (actually it was 67/6 which is 11.133, but have you ever seen .133 of a room?). So, 12 rooms (all prepared in exactly the same way) for one day, representing a total of 24 personnel. All candidates had to be carefully scheduled and informed of their test time. Additional staff were then required to check that candidates were present and in place at the correct time. The total administration time for this test was in the region of 500 hours (remember that all personnel were trained and on the day worked for 6 hours and that all results had to be collated, analysed and reported). This represents a great deal of time and, of course expense.

2. The Test Taker

By systematically defining the test takers, we can genuinely take them into account when designing tests. One example of this is that we would make decisions related to the appropriacy of reading texts for the intended test population based on a broad range of parameters, or that we would tailor our expectations (in terms of expected linguistic output for example) again based on a broader understanding of the population. Too often the test developer bases important decision on their perception of a test population rather than on evidence. Where there is evidence that the population is heterogeneous with regard to a number of the characteristics described here, we have a problem. An example of this is where the age and background of the population taking a test varies greatly. Here, the test developer struggles to come up with tasks that are likely to result in the best performances from all candidates, and is one reason why large-scale international tests (such as TOEFL and Cambridge ESOL Main Suite examinations) are considered by many to be very bland.

The other relevance of understanding the test taker is to allow for the provision of accommodations (i.e. special circumstances) for students with disabilities. Nowadays, there is extensive legislation covering this area (certainly across the EU, in Asia and in the Americas). A good example of what an examining board is expected to take into account can be found on the Cambridge ESOL website (though all reputable examination boards will publish lists similar to that reproduced below). The major categories of accommodation (Cambridge ESOL refers to them as 'special arrangements') available for test takers in the Cambridge ESOL examinations were listed by Taylor (2003) as:

- Braille Versions
- Enlarged Print Versions
- Hearing-impaired (lip-reading) Versions
- Special Needs Listening Test Versions
- Separate marking for candidates with Specific Learning Difficulties
- Exemption from Listening or Speaking components
- Additional time, provision of a reader and colour paper/overlay for dyslexic test takers

To give an example of the complexity of the situation (outside of the fact that a single application may include a request for accommodations in a number of areas) it is useful to look at the range of accommodations offered just for the speaking papers, where there are usually two learners working together and sometimes three learners. This can cause some problems with the provision of accommodations, as the welfare of all test takers must be taken into account if the test is to be fair to all participants.

Hearing Difficulties

- *extra time* (if it takes longer than usual to say things or to understand what people are saying)
- *a partner who is not doing the examination* (e.g. it is easier for the hearing impaired test taker to lipread what the partner is saying if that partner can focus on clearly articulating, or lip speaking, each word rather than on their own performance)
- *no partner* (i.e. in those parts of the test which usually ask both candidates to talk to each other, the test taker may talk to the examiner instead – though this option is only available for the Main Suite tests).
- *Note: signing is not allowed (as this is considered a different communication skill)*

Visual Difficulties

- extra time (where a test taker takes longer than usual to read any exam material or decide what they want to say)
- *a partner who is not doing the examination* (an arrangement designed to eliminate any bias towards or against the partner of a candidate qualifying for special arrangements)
- *no partner* (like the arrangement for test takers with hearing difficulties, this applies to those parts of the test which usually ask both candidates to talk to each other. Again, the test taker may talk to the examiner instead).
- *adapted visual material* (e.g. Braille versions of the task input material)

Short-term Difficulties

In the case of short-term difficulties, such as minor illnesses or injuries, test centres are encouraged to take a supportive attitude, for example by bringing forward or delaying the speaking test paper where possible. Taylor (2003: 2) reports that from a total testing population of close to four hundred thousand (my estimate) for the Cambridge ESOL Upper Main Suite examinations (FCE, CAE & CPE) in 2001, the number requesting accommodations for the Speaking & Listening papers was just 11.

TEST TAKER CHARACTERISTICS	
Physical/Physiological	
Short term ailments	*Toothache, cold etc.*
Longer term disabilities	*Speaking, hearing, vision (e.g., dyslexia)*
Age	Suitability of materials, topics etc. Demands of tasks (time, cognitive load etc)
Sex	Suitability of materials, topics etc.
Psychological	
Memory	Related to task design, also to physical characteristics
Personality	Related in speaking primarily to task format (e.g. number of participants in an event – solo, pair, group, etc. can impact on how shy learners will perform)
Cognitive Style	This refers to the way individuals think, perceive and remember information, or their preferred approach to using such information to solve problems (if a task is primarily based on one aspect of input such as a table of information, this may negatively affect some candidates)
Affective Schemata	How the candidate reacts to a task. Can be addressed by the developer through carefully controlled task purpose (even a sensitive topic can be addressed if the candidate is given a reasonable purpose – e.g. allowing a candidate to personalise a topic can help them negate many adverse affects) and/or topic (all examination boards have taboo lists – i.e. list of topics to avoid, such as death, smoking etc.)
Concentration	Related to age and also to length and amount of input
Motivation	Among other things this can be related to task topic or to task/test purpose
Emotional state	An example of an unpredictable variable. Difficult to deal with, though may be approached from the same perspective as Motivation or Affective Schemata.
Experiential	
Education	This can be formal or informal and may have taken place in a context where the target language was either the principal or secondary language
Examination Preparedness	Can relate either to a course of study designed for this specific examination, examinations of similar design or importance, or to examinations in general.
Examination Experience	Again can relate to this specific examination, examinations of similar design or importance, or to examinations in general.
Communication Experience	Can relate to any of the above, e.g. where communication experience is based only in classroom interactions or where the candidate has lived for some time in the target language community and engaged in 'real' communication in that language.
TL-Country Residence	Can relate to Education (i.e. place of education) or to Communication Experience (e.g. as a foreign or second language)

Table 1 Test-Taker Characteristics

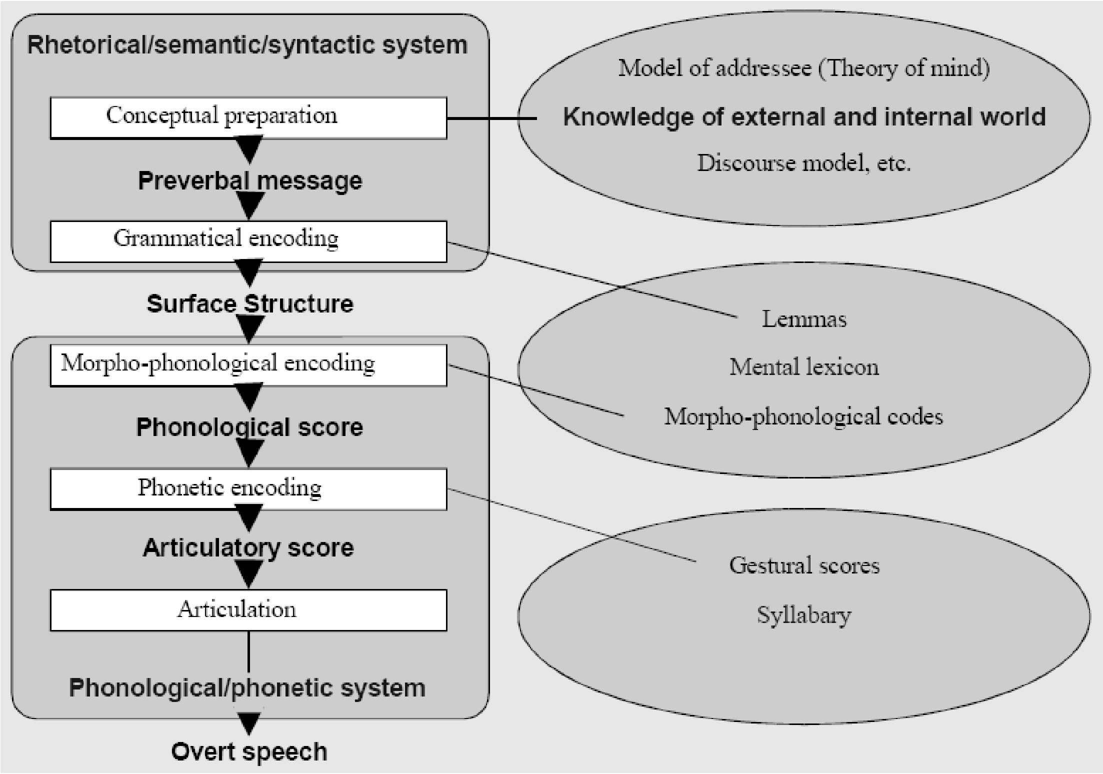
3. The Theory-Based or Cognitive Perspective

3.1. Language Processes

In the 1970s, the area of psycholinguistics was most obviously associated with studies in spoken language understanding and processing. At that time, there were two commonly held views: first that processing is sequential with each component being autonomous in its operations; and second that processing is a more flexibly structured system (Fodor et al., 1974; Marslen-Wilson and Tyler, 1980; Marslen-Wilson et al., 1978; Tyler and Marslen-Wilson, 1977). However, the primary concern for psycholinguists was in fact how spoken language relates to underlying linguistic systems.

Levelt (1989) was the first to model the processing system that underlies speech production (see Figure 1). His model (or blueprint as he called it) illustrates the organization of the speech process, from the constraints on conversational appropriateness to articulation and self-monitoring. This was a more comprehensive system than previously theorised and remains the predominant theory today. Seeing the speaker as an information processor, Levelt proposes a blueprint in which message generation, grammatical encoding, phonological encoding, and articulation are seen as relatively autonomous processors. Two points are however, made clear. First, though Levelt 's model stops largely at the point of utterance, he devotes an entire chapter on the speaker as an interlocutor in natural conversation (Chapter 2). Here he describes at length the three essential aspects of conversation in which the speaker is a participant and interlocutor: it is highly contextualized, has spatio-temporal setting, and is purposeful. Secondly, though the model may seem complex at first, the basic mechanisms of speech processing are conceptualized in a fairly uncomplicated manner: we produce speech by first conceptualizing the message, then formulating its language representation (encoding) and finally articulating it. With reference to speech perception, speech is first perceived by an acoustic-phonetic processor, then linguistic encoding in the speech comprehension system (the parser), and it is finally interpreted by the conceptualizer.

Figure 1 A Blueprint of the Speaker (Levelt, 1989; 1999)



Levelt's work in terms of the blueprint/framework influenced other works and has been used or referenced in more recent works on research in speaking (Bortfeld et al., 2001; Dornyei and Kormos, 1998; Ellis, 2003; Hughes, 2002; Ortega, 1999; Weir, 2005).

The Levelt blueprint forms the foundation for theory-based validity/internal processing component of the framework for validating a speaking test (Weir, 2005). This aspect of the validity framework is essential, not just for the purpose of validation but also for a better understanding of the processes or operations that test takers utilize when attempting the test task; only through such data can we make decisions about these operations in relation to the elements we include in the test task or context validity. See Table 2 for an overview of how Levelt's work impacts on speaking test development.

COGNITIVE VALIDITY	
COGNITIVE PROCESSES – based on Levelt (1989)	
Conceptualiser	conceiving an intention, selecting relevant information to be expressed to realize this purpose, ordering information for expression, keeping track of what was said before; paying constant attention to what is heard and own production, drawing on procedural and declarative knowledge. Speaker will monitor messages before they are sent into the formulator.
Pre verbal message	product of the conceptualisation stage
Linguistic formulator	includes grammatical encoding and phonological encoding which accesses lexical form
Phonetic plan	an internal representation of how the planned utterance should be articulated; internal speech
Articulator	the execution of the phonetic plan by the musculature of the respiratory, the laryngeal and the supralaryngeal systems
Overt speech	
Audition	understand what is being said by others or self, i.e. interpret speech sounds as meaningful words and sentences
Speech comprehension	access to various executive resources e.g. lexicon, syntactic parser, background knowledge. A representation is formed of the speech in terms of its phonological, morphological, syntactic and semantic composition. Applies to both internal and external overt speech.
MONITORING	both of internal and external speech can be constantly in operation though sometimes this filter is switched off. The system through which internal resources are tapped in response to demands of executive processing.
COGNITIVE RESOURCES	
Content knowledge	
Internal	The test-taker's prior knowledge of topical or cultural content (background knowledge)
External	Knowledge provided in the task
Language knowledge – all references to Buck (2001)	
Grammatical	literal semantic level: includes phonemes, stress, intonation, spoken vocabulary, spoken syntax
Discoursal	related to longer utterances or interactive discourse between two or more speakers: includes knowledge of discourse features (cohesion foregrounding, rhetorical schemata and story grammars) and knowledge of the structure of unplanned discourse
Functional	function or illocutionary force of an utterance or longer text + interpreting the intended meaning: includes understanding whether utterances are intended to convey ideas, manipulate, learn or are for creative expression, as well as understanding indirect speech acts and pragmatic implications
Sociolinguistic	the language of particular socio-cultural settings + interpreting utterances in terms of the context of situation: includes knowledge of appropriate linguistic forms and conventions characteristic of particular sociolinguistic groups, and the implications of their use, or non-use, such as slang, idiomatic expressions, dialects, cultural references, figures of speech, levels of formality and registers

Table 2 Cognitive Validity & Levelt

2.2 Language Knowledge

In Part 2 we discussed test validation. Here, we saw that language knowledge refers to assumptions on the part of the test developer of how test takers' language can be most clearly defined. It is important to note that the efforts of testers such as Bachman (1990) to define language ability (more accurately communicative language ability) stemmed from the desire to base tests on an operationalised model so that the underlying construct could be made clear. Rather than repeat the argument made in that part of the course, it is best to focus on the central importance of having an explicit understanding (or theory) set out before continuing with the development cycle. This advice goes for the testing of any skill or ability.

While it is important to identify the aspects of language to be examined, there is some considerable evidence to suggest that it is not possible to predict language use in task performance at the microlinguistic level (e.g. grammar or lexicon). Researchers and language testers have instead focused on more macro linguistic descriptions of language. Bygate (1987Chapter 4, pp. 23-41) provides a useful categorisation of the type of operations involved in any communicative interaction. This categorisation has been adapted by Weir (1993: 34) in his checklist of operations:

Routine Skills	**Improvisation Skills**	**Microlinguistic elements**
Informational	*Negotiation of Meaning*	
expository	*Management of Interaction*	
evaluative		
Interactional		

It should be noted that this checklist represents an attempt to gather together a set of criteria by which the language of the test may be predicted. In identifying the operations involved in the performance of a particular task, we are essentially defining the construct that we are attempting to examine through that performance.

Informational	Interactional	Managing Interaction
Providing personal information	Challenging	Initiating
Providing non-personal information	(Dis)agreeing	Changing
Elaborating	Justifying/Providing support	Reciprocating
Expressing opinions	Qualifying	Deciding
Justifying opinions	Asking for opinions	Terminating
Comparing	Persuading	
Complaining	Asking for information	
Speculating	Conversational repair	
Analysing	Negotiating meaning	
Making excuses		
Explaining		
Narrating		
Paraphrasing		
Summarising		
Suggesting		
Expressing preferences		

Table 3 Discourse Functions (from O'Sullivan, Weir & Saville 2002)

O'Sullivan, Weir & Saville (2002) presented the results from such a validation project, carried out at Cambridge ESOL with their Main Suite examinations in mind. The project involved an attempt to predict the language functions elicited by a set of tasks (in the FCE examination) using a set of checklists developed from the above categorisation. The study suggests that it is possible to perform such a prediction, making it feasible to use a functions checklist in the writing of tasks and later in the review of actual performance. Surprisingly, this has not been attempted to date, even though (as the authors point out) it would appear to be of central importance that our predicted outcome or response will match the actual. The checklists used in the study are presented in Table 3.

3.3 Background Knowledge

Background knowledge has not been considered a major factor in the testing of speaking. This is because of the task types typically used, where knowledge is presented in the task input in order to ensure that all language samples elicited will be comparable. While there are some examples of so-called 'free conversation' type oral examinations, the general consensus nowadays is that the inherent inequality of the test event makes true *conversation* (where all interlocutors have equal control over the discourse) impossible, see van Lier (1989).

4. The Context Perspective

In the Table 3 we can see how each parameter can be operationalised. This way of looking at the different parameters is actually very useful as it allows the test developer a simple, yet systematic framework for describing each test task (so it can be used to build a specification for the test). It also suggests a methodology for systematically evaluating a test task, e.g. think about a test task you are familiar with (IELTS Writing Task A; TOEFL Reading Paper etc.) now ask questions of the task based on each parameter.

I have also used this format to review a series of tests (in any of the four skills) that are claimed to be progressing through levels of ability. To do this, simply create a single table with a column for each test to be reviewed. When you complete the information asked for each test you can then make observations about the relative level and/or complexity of each test in comparison to the others. This is particularly relevant for programmes where there a number of progressive levels.

CONTEXT VALIDITY	
Settings: Task	
Purpose	The requirements of the task. Allow candidates to choose the most appropriate strategies and determine what information they are to target in the text in comprehension activities and to activate in productive tasks. Facilitates ***goal setting*** and ***monitoring.***
Response format	How candidates are expected to respond to the task (e.g. MCQ as opposed to short answers). Different formats can impact on performance.
Known criteria	Letting candidates know how their performance will be assessed. Means informing them about rating criteria beforehand (e.g. rating scale available on WEB page).
Weighting	Goal setting can be affected if candidates are informed of differential weighting of tasks before test performance begins.
Order of Items	Usually in speaking tests this is set, not so in writing tests.
Time constraints	This can relate either to pre-performance (e.g. planning time), or during performance (e.g. response time)
Intended operations	A broad outline of the language operations required in responding to the task. May be seen as redundant as a detailed list is required in the following section.
Demands: Task *[note: this relates to the language of the INPUT and of the EXPECTED OUTPUT]*	
Channel	In terms of input this can be written, visual (photo, artwork, etc), graphical (charts, tables, etc.) or aural (input from examiner, recorded medium, etc). Output depends on the ability being tested.
Discourse Mode	Includes the categories of genre, rhetorical task and patterns of exposition
Text Length	Amount of input/output
Writer/speaker relationship	Setting up different relationships can impact on performance (e.g. responding to known superior such as a boss will not result in the same language as when responding to a peer).
Nature of Information	The degree of abstractness. Research suggests that more concrete topics/inputs are less difficult to respond to that more abstract ones.
Topic familiarity	Greater topic familiarity tends to result in superior performance. This is an issue in the testing of all sub-skills
Linguistic	
Lexical Range	these relate to the language of the input (usually expected to be set at a level below that of the expected output) and to the language of the expected output. Described in terms of a curriculum document or a language framework such as the CEFR.
Structural Range	
Functional Range	
Interlocutor	
Speech Rate	Output expected to reflect that of L1 norms. Input may be adjusted depending on level of candidature. However, there is a danger of distorting the natural rhythm of the language, and thus introducing a significant source of construct-irrelevant variance.
Variety of Accent	Can be dictated by the construct definition (e.g. where a range of accent types is described) and/or by the context (e.g. where a particular variety is dominant in a teaching situation).
Acquaintanceship	There is evidence that performance improves when candidates interact with a friend (though this may be culturally based).
Number	Related to candidate characteristics – evidence that candidates with different personality profiles will perform differently when interacting with different numbers of people.
Gender	Evidence that candidates tend to perform better when interviewed by a woman (again can be culturally based), and that the gender of one's interlocutor in general can impact on performance.

Table 3 Context Validity

4.1 Language Elicitation Tasks for Speaking

The following set of task types represents an effort to somehow collapse the vast range of test tasks that have been used in tests of spoken language ability. This is not meant to be a complete set, but instead may be used as a guide or framework in which tasks may be ordered. Unlike the previous sections, it may be seen from this list that it is not terribly difficult to create a test which elicits a sample of a learner's spoken language. However, as we will see later in this section, this is only the beginning. The sample must be rated (or given some kind of score) so that the performance is made 'usable', in other words, stakeholders demand that any test results should be reported in a way that they can understand and *use*.

Task Type	Description	Advantage(s)	Disadvantage(s)
1. Reading Aloud	Student normally asked to silently read a text then to read it aloud to the examiner	All students must read the same text so a similar level of performance is expected, makes for ease of comparison Language can be easily controlled	There are significant differences in native speaker performance. Interference between reading and speaking skills. In no way valid, while remaining open to unreliability (subjective assessment used). Seen as unacceptable in most books.
2. Mimicry	Students are asked to repeat a series of sentences after the examiner. Results recorded and analysed	Can be performed in a language laboratory with a large number of students at one time. Students expected to perform equally as input is same for all. Language easily controlled. Research shows error type similar to 'free' talking.	Difficult to interpret, and therefore to score, the results. Not authentic. Not communicative. Evaluates other skills such as short term memory and listening. Severe 'Backwash' effect.
3. Conversational Exchanges	Students are given a series of situations (read or heard) from which they are expected to make sentences using particular patterns. Models of the expected language may or may not be first given, this changes the nature of the task.	Suitable for use with a large number of students, for example in the language laboratory. Language is controlled, so comparison is possible and reliability is likely to be high. Content validity in that the language tested will be directly related to that studied in class by the students.	No authentic interaction, therefore the test is in no way communicative. Reading or listening skills will interfere with the student's ability to respond to the stimulus. At best it tests a student's ability to reproduce the chosen patterns under extremely limited conditions.

Task Type	Description	Advantage(s)	Disadvantage(s)
4. Oral Presentation (Verbal Essay)	Student asked to speak, without preparation, (usually 'live' though occasionally directly onto tape) for a set time (e.g. 3 minutes) on one or more specified general topics. In an alternative version some time may be allowed for preparation (e.g. 30 seconds or 1 minute).	As students must speak at length a wide variety of criteria may be included in any evaluation (inc. fluency)	Topic may not interest student. Not authentic to 'real' life. Offering a choice of topics makes comparison difficult More open ended topics and the lack of preparation time may mean that performance depends on the extent of the learners' background (non-linguistic) knowledge. Use of tape recorder may add to the stress of the student.
(Prepared monologue)	Similar to the Verbal Essay but the student is given time to prepare	Easy to prepare and to administer Gives the 'appearance' of a communicative task.	Likely native speaker differences make it an unreliable and invalid procedure. Students likely to memorise text. Unless same monologue is given to all students, results not comparable Knowledge of or interest in the topic will affect performance With insufficient preparation time students' knowledge may be tested and not their language.
5. Information Transfer (Description of Picture Sequence)	Students take a series of pictures and try to tell the story in a predetermined tense (e.g. the past) having had some time to study the pictures	Clear task. If cultural/educational bias is avoided in the pictures no contamination of the measurement takes place. Elicits extended sample of connected speech. Examines students' ability to use particular grammatical forms. Students exposed to same prompts, so performance comparisons valid.	Limited authenticity. Tells little of students' ability to interact orally. Poor picture quality can affect student performance. Reliability of scores may be affected by differences in interpretation of the pictures.
(Questions on a single Picture)	Examiner asks student several questions about the content of a particular picture, having first given them time to study it.	Can offer authentic materials to the student, especially where the content in geared to the interest of the student	Student can only respond to the questions asked. Picture must be clear and unambiguous. If large scale difficulties of comparability and of test security arise.
(Alternative Visual Stimuli)	Where 'real' objects are used instead of pictures as stimuli	Similar advantages to the student as with a picture elicitation task, while adding a touch of greater reality.	Similar disadvantages to using a picture. A knowledge of the object in question may interfere with the language produced.

Task Type	Description	Advantage(s)	Disadvantage(s)
6. Interaction Tasks (Information Gap: Student - Student)	Usually done in pairs, each person is given part of the total information, they must then work together to complete a task.	When students are free to select their partner this can be one of the most effective communicative tests tasks. Generates a wide variety of criteria on which rating is dependent Highly interactive.	One participant may dominate. Large proficiency differences may affect performance One student may be more interested in the task. Presents one situation of language use. Practical problems include time, administration difficulties, and maintaining test security.
(Information Gap: Student - Examiner)	As above, but with a student who is missing information required to complete a task and must request it from the examiner, who acts as the interlocutor.	Interlocutor may act in a similar way with all candidates, making performance comparison valid.	Can be very daunting for the student. Examiner may be assessing own performance in addition to that of the student. [examiner may not always interact the same way with all students]
Role Play (open)	Student expected to play one of the roles in an interaction possible in 'real' language use. Can be Student - Student, or Examiner - Student.	Face and content validity in a variety of situations. May be a reliable way of observing and measuring a students ability to perform in given situations	'Histrionic' students may score higher than more introverted ones. Role familiarity may affect performance. Students sometimes use 'reporting' language instead of adopting the role. When large scale, different role plays are required, causing problems with comparability and security.
Role Play (guided)	Examiner (or volunteer) takes a fixed (scripted) role in a roleplay situation. Student responds to these prompts.	Examiner has great control over the language elicited. 'Situation' may be controlled to reflect present testing requirements or objectives.	Using different topics may increase user-friendliness of task but will make result comparison impossible. Does not allow for genuine interaction/topic expansion therefore not really a communicative test.
7. Interview (free)	No predetermined procedure, conversation "unfolds in an unstructured fashion."	High face and content validity.	Performance varies due to different topics and due to differences in the way the interview is conducted. Time consuming and difficult to administer
(Structured)	Normally a set of procedures is used to elicit performance, that is there are a series of questions and/or prompts to guide the interviewer through the interview.	Greater possibility of all sts being asked the same questions, therefore comparisons more valid. High degree of content and face validity. High inter and intra-rater reliability with training.	Limited range of situations covered. Examiners may not always stick to the predetermined questions.

Task Type	Description	Advantage(s)	Disadvantage(s)
8. Discussion (Student - Student)	In pairs or groups students are asked to discuss a topic, make plans, etc.	Good face and content validity. Communicative. Topics may be teacher or student determined.	Performance of one strong individual may dominate others, so it needs careful matching. Possible interference from conversation management and discussion skills. A relatively large number of students may be tested at one time
(Student - Examiner)	Examiner determines topic (though it could be determined in cooperation with the student) then guides the discussion.	High face validity if student has a role in the topic choice. Format may encourage the examiner to expand the conversation in chosen directions.	Unless the same topic is employed with all students the resulting scores will not be comparable. Even if the same topic is used there is little likelihood that all students will produce the same language. Difficult to attain high reliability with just one examiner/scorer.

4.2 Speaking Test Formats

In addition to presenting a set of tasks, as in the previous sections, we must also think about the different formats used in tests of spoken language. Of course there is no clear line between the two sets of lists presented here, as there are some tasks which are associated with particular formats and others which could well be used (or adapted for use) in a number of formats. For obvious reasons the following list of formats is not complete and represents an effort to outline some of the principal formats used, in addition to mentioning some lesser known ones.

Format	Description	Advantage(s)	Disadvantage(s)
1. Candidate Monologue	Candidate performs set task (typically task 4 above, though tasks 1, 2 or 3 could also be used) either 'live' or recorded.	As listed for tasks 1 to 4.	As listed for tasks 1 to 4.
2. Interview	Candidate is examined either alone or as a member of a pair. The interviewer asks a series of predetermined questions of each candidate. Known as the Oral Proficiency Interview (OPI) and is the basis for the CAMBRIDGE ESOL spoken components.	As listed for task 7.	As listed for task 7.

Format	Description	Advantage(s)	Disadvantage(s)
3. Recorded Stimuli	Questions first tape-recorded, then students listen and respond (response also recorded). Known as the Simulated Oral Proficiency Interview (SOPI) and is used in the Test of Spoken English (TSE).	Uniform results expected, so can be used for comparison Suitable for use in a language laboratory Relatively easy to score, and reliable.	Inflexible, no possibility of expansion or follow up on students' answers. Not authentic, no verbal or non-verbal feedback possible. Can be very time-consuming for the examiner.
4. Alternative Formats (self evaluation)	The student is asked to evaluate own language performance/ ability, using a predetermined scale.	Easy for the teacher to set once the scale has been settled on Useful to encourage student self evaluation outside of the testing situation	Certainly in the early stages of use it is not reliable. Can be culturally influenced, therefore is not suitable for a mixed-culture group.
(teacher evaluation)	Teacher continually assesses student ability and performance during the term.	With (almost) daily contact the teacher is in a unique position to longitudinally assess the student. As the final score awarded is based on a large number of evaluations it will probably be valid and reliable.	Open to interference from the student/ teacher interpersonal relationship. Only really useful when combined with another test result. Use is limited to course evaluation. Should not be used as a placement test as variables such as student attendance will interfere.
(peer evaluation: interview)	In groups of three or four students take turns as interviewer, observer and interviewee, during which they are asked to score the interviewee's performance on a predetermined scale.	Large classes can be accommodated in 30 to 40 minutes (each interview lasts approx. 10 minutes). A limited number of variations makes it replicable with the same group. Removes some student test apprehension. Research data shows a high rate of agreement among interviewer and observer raters	Teacher has limited 'control' over each interview. Scoring can be influenced by factors other than language ability, such as the inter-student relationships in the group. May be more effective with older or more highly motivated students.
(peer evaluation: group / pair work or roleplay)	As with examiner monitored tasks except that here the evaluation is performed either by individuals in the pair/group or by other student observers.	Similar advantages to the peer evaluated interviews. Where pairs / groups perform individually with remaining sts acting as raters the reliability will tend to be high. If individually tested examiner may observe performances to provide additional score	Similar disadvantages to the peer evaluated interviews and to the examiner evaluated group / pair work or roleplay tasks. Asking individuals to rate each other when they are all equally engaged in the task may be beyond the scope of most younger or lower level students.

It is quite common that a test will be made up of series of tasks, each involving a different degree of cognitive load (for example a personal information exchange task is relatively light as the response is already known so the candidate can focus on the language, while a decision making task is heavier as the answer is not known — there is none! — so there is less opportunity to focus on the language), and expected response (in terms of the performance conditions). An example of this is the Cambridge ESOL First Certificate in English (FCE) spoken language test (Paper 5). In this test there are two candidates (C_1 & C_2), an interlocutor/examiner/facilitator (IEF) and an observer (O). The test begins with monologues from each of the candidates, followed by an interactive task between the candidates and finishing in a three-way interaction task (C_1, C_2 & IEF).

4.2.1. Test Example

The above descriptions indicate that tests of speaking generally follow one of two main methods, live or recorded. This can consist of an interview (where a candidate or candidates communicate with an examiner in a one-to-one interaction), a pair-work activity (or activities), or of a combination of these formats. The former of these are often referred to as Oral Proficiency Interviews (OPIs) though in North America this is typically associated with the Foreign Services Institute (FSI) test.

In the following two sections I will present brief case studies on these two formats. The first test is the Cambridge ESOL FCE speaking paper.

The FCE Speaking Paper

A good example of a set of tests which attempt to include all of the above (in addition to the inclusion of a student monologue) is the Cambridge ESOL Main Suite battery of examinations. These tests are aligned with the ALTE framework and between them are expected to cover the range proficiency (see Figure 2)

Figure 2 The Cambridge/ALTE Five-Level System

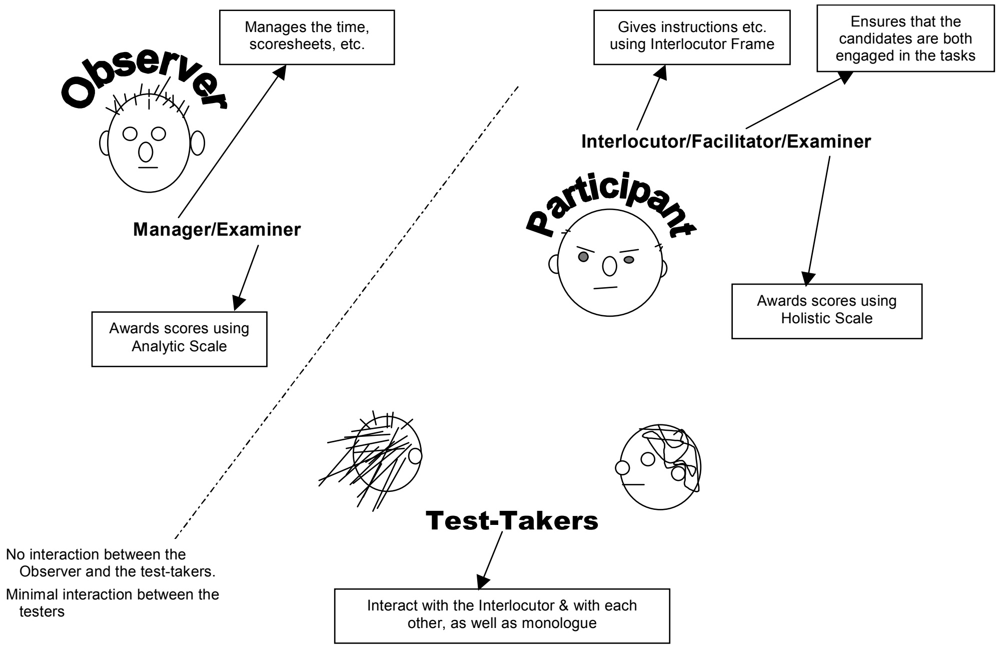
ALTE Level 1 ***Waystage User*** **CAMBRIDGE Level 1** ***KET***	**ALTE Level 2** ***Threshold User*** **CAMBRIDGE Level 2** ***PET***	**ALTE Level 3** ***Independent User*** **CAMBRIDGE Level 3** ***FCE***	**ALTE Level 4** ***Competent User*** **CAMBRIDGE Level 4** ***CAE***	**ALTE Level 5** ***Good User*** **CAMBRIDGE Level 5** ***CPE***
BASIC	**INTERMEDIATE**		**ADVANCED**	

The FCE Speaking Test Format

The FCE Paper 5 is a direct test of speaking ability using a paired format of two test-takers and two testers, with, in exceptional cases, the possibility of a group of three test-takers. Saville and Hargreaves (1999: 44) justify this format on the grounds of stakeholder feedback, greater reliability and fairness from having two ratings of the performance, the broadening of the interaction-types available, and its potential for positive washback through encouraging the use of pair-work activities in the language learning classroom. While the test format has been criticised, particularly by Foot (1999), for its failure to acknowledge the potential difficulties relating to dyad composition (different language level, sex, personality etc.), this criticism has not been based on any empirical evidence – though the questions raised are important, and will be directly addressed in this study.

The figure below (Figure 3) represents the physical format of the test, and includes reference to the various roles undertaken by the testers during a test event.

Figure 3 The FCE Paper 5 (Speaking) Format


Observer
Manages the time, scoresheets, etc.
Manager/Examiner
Awards scores using Analytic Scale
Gives instructions etc. using Interlocutor Frame
Ensures that the candidates are both engaged in the tasks
Interlocutor/Facilitator/Examiner
Participant
Awards scores using Holistic Scale
No interaction between the Observer and the test-takers.
Minimal interaction between the testers
Test-Takers
Interact with the Interlocutor & with each other, as well as monologue

The test content

The test consists of a number of tasks, each designed to elicit a different sample of language functions, so that each candidate has an opportunity to demonstrate the range of his/her ability. See Figure 4 for an outline of the four tasks.

Figure 4 Format of the Main Suite Speaking Tests

1	**Interviewer – Candidate**	**Interview** Verbal Questions
2	**Candidate**	**Monologue** Individual Long turn
3	**Candidate – Candidate**	**Collaborative task** Visual stimulus Verbal instructions
4	**Interviewer – Candidate – Candidate**	**Long turns and discussion** Written stimulus Verbal Questions

The test-takers

As mentioned above, over 250,000 test-takers from over 150 countries throughout the world participate in the FCE every year. Approximately 75% of these test-takers are aged under 25, with the average being 23 years, though in some countries, such as Greece, the average is lower. Most test-takers are female, and are students, though these figures differ from country to country. In addition, approximately 80% of the test-takers will have undertaken a course of study designed specifically for the FCE (CAMBRIDGE ESOL 1997).

During the test, these test-takers are expected to perform a set of four tasks which involve a series of interactions with the other participants. These interactions will be discussed in the section related to the test tasks.

The testers

The FCE format involves two testers, one who participates directly in the interaction, and another who remains a neutral observer. Within these roles each tester undertakes a number of functions. Similarly, the test-takers are expected to undertake different roles, in terms of the different interaction types demanded of them by the different tasks.

The Tester as Participant

The role of the tester (the person who will interact with the test-takers) is quite complex as it varies with each task during the test. These roles are:

Interlocutor: The interlocutor interacts directly with each test-taker at all stages of the test, particularly in the interview stage. The interlocutor's role is related to the fabric of the test, in that (s)he is expected to follow a predetermined 'interlocutor-frame' which is scripted or controlled.

Facilitator: The facilitator must accommodate the interaction between two test-takers, encouraging both during the monologic and dialogic stages. The role of facilitator differs from that of interlocutor in that it is in this role that the tester must exercise a degree of spontaneity independent of the 'interlocutor-frame' in order to ensure that each interaction is engaged in as equally as possible by the test takers.

Examiner: The examiner must award a score, using a Holistic scale (described below) to each test-taker based on their performance.

The Tester as Observer

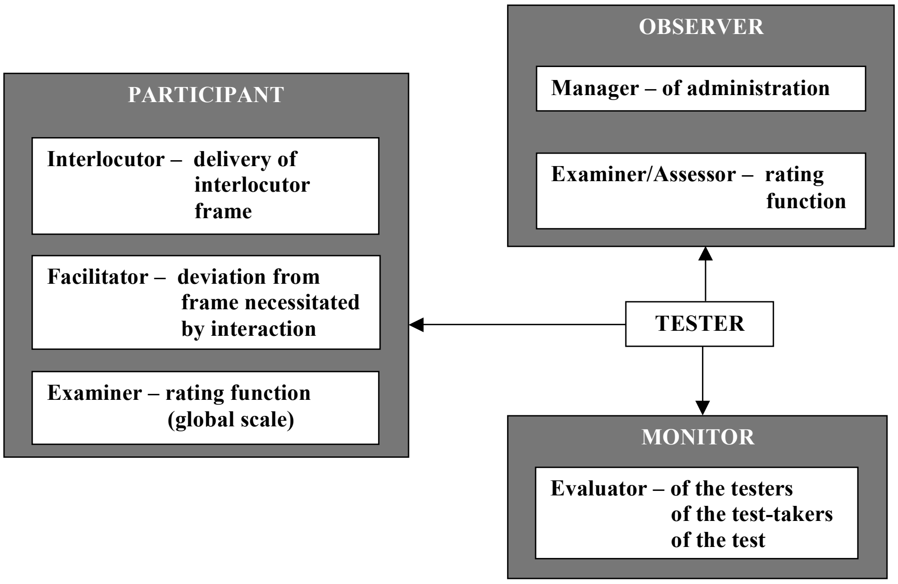
The role of the tester-as-observer (the person who does not interact with the test-takers) is less complex though this person also takes on a number of roles during the test. These roles are:

Manager: The observer-as-manager's role is related to the management of the administration of the test, principally in terms of ensuring that the score sheets have been correctly completed and scores have been entered correctly.

Examiner: As an examiner, the observer has far more time in which to evaluate the test-takers' performances in relation to a number of criteria. Therefore, he or she must award a score using an Analytic scale (comprised of four equally weighted criteria and described below) to each test-taker based on their performance.

In addition to these roles, individual tests are selected for monitoring – both of the test and tasks, and for tester performance. This monitoring is typically done by a senior examiner called a Team Leader. While this additional role is recognised here, it is not relevant to this study; although some random monitoring was done on the tests being reported here, it will not feature in this study as the monitor neither plays a part in the administration of the test, nor interacts with either the testers or the test-takers. Figure 5 represents an overview of the tester roles within the FCE.

Figure 5 The role of the Tester in the FCE Testing Event


OBSERVER
Manager – of administration
Examiner/Assessor – rating function
PARTICIPANT
Interlocutor – delivery of interlocutor frame
Facilitator – deviation from frame necessitated by interaction
Examiner – rating function (global scale)
TESTER
MONITOR
Evaluator – of the testers of the test-takers of the test

4.2.2. The Simulated Oral Proficiency Interview [SOPI]

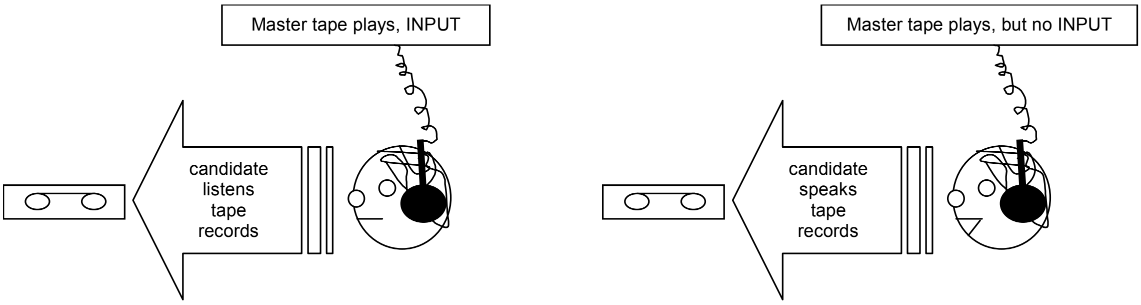
The Simulated Oral Proficiency Interview (SOPI) is a tape-mediated test of speaking ability – as opposed to the Oral Proficiency Interview (OPI) which is a live test. In other words, it is performed by a group of test-takers in a language lab. One advantage of this type of test is that it can cope with quite large numbers of students while providing standardised input at all times. A typical SOPI will contain a series of different tasks, designed to access different aspects or dimensions of a candidate's speaking ability. Performance on all tasks is typically scored using an Analytic scale (though Holistic scores are common), the final score is then averaged from all scores awarded. Another advantage of this method is that the performances can be rated directly from the tapes by trained raters working independently of each other.

A typical example of this type of test is the Graduating Students Language Proficiency Assessment (GSLPA) taken by all graduating student at the Hong Kong Polytechnic University. This test consists of five tasks:

1. Listen and report – based on a five minute radio interview (report to friend)
2. Interview – series of 4 questions based on job advertisement
3. Short Presentation – based on given information (to work colleagues)
4. Telephone Message – listen to request, reply to request (work related – to colleague)
5. Social Context – respond to request for some specific information about Hong Kong.

The test works by playing the master tape to each of the booths in the language lab. Each candidate listens to the tape and responds when cued – the response times are built in to the master tape. Meanwhile individual tapes are recording everything that happens in each booth (the master tape and the responses) so the examiner is left with evidence of the input and of the output. The resulting tapes are then multiple-rated (by trained raters).

Figure 6 The SOPI Format


Master tape plays, INPUT
candidate listens tape records
Master tape plays, but no INPUT
candidate speaks tape records

This method has been criticised as it results in monologic discourse only (though you can see how the designers have tried to build in a context and an audience to limit this effect), and for being unnatural (or inauthentic). Since the format was first proposed (in the early 1980s) the test has moved with the times with more recent VOPIs (Video) and COPIs (Computer) proposed.

While there have been some studies that attempted to explore the differences between the two formats (SOPI/OPI), these have not been definitive in their outcomes (O'Loughlin, 2001) – though it appears to be clear that raters tend to he harsher when rating taped performances.

5. The Scoring Perspective

Finally, we will take a look at how spoken performance is assessed. This is a central aspect of Scoring Validity.

In the past, the emphasis on performance test (writing or speaking) reliability tended to focus on reliability (typically inter-rater reliability). This emphasis, while useful, seriously limits our overall understanding of how every aspect of the process of turning a test performance into a score or grade is important to the overall validity of inferences drawn from that score or grade. We therefore see (in Table B6) that we should pay attention to every step of the process. This is not to ignore the importance of the measurement qualities of a test. It is still vitally important that any test meet the highest possible standards, so we would still expect to investigate the inter- and intra-rater reliability of any productive language test.

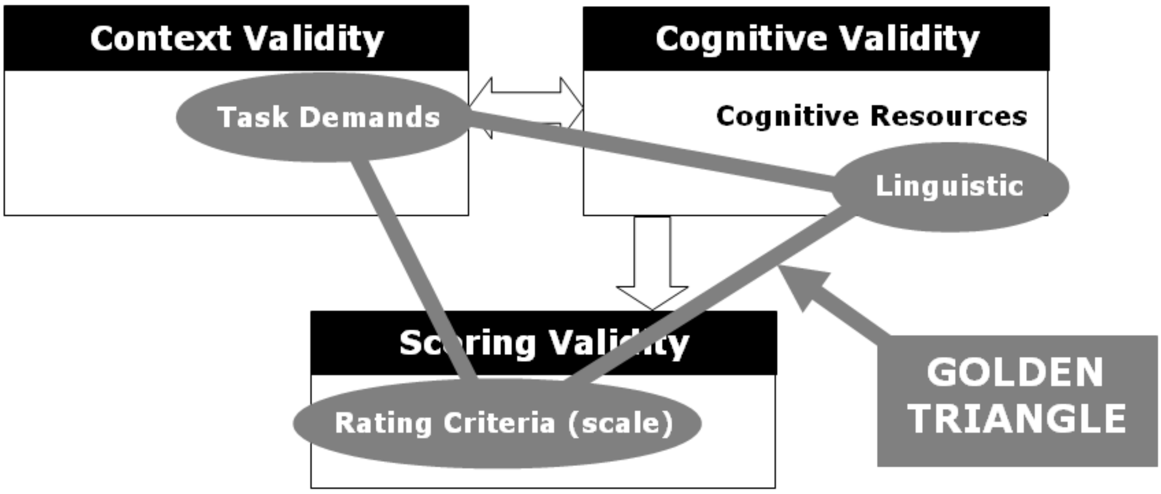
SCORING VALIDITY	
Criteria/Rating Scale	The criteria must be based on the theory of language (Language Knowledge) outlined in the Theory Based Validity section and reflected again in the Demands: Task section of Context Validity. They should also reflect 'actual' language production for the task or tasks included in the examination.
Rating Procedures	
Training	There are a number of different approaches to training, and there is evidence that training improves harshness, consistency and ability to stay on standard.
Standardisation	As part of any training regime, raters must internalise the criterion level (e.g. pass/fail boundary) and this should be checked using a standardisation procedure (or test if you like).
Conditions	Attempts should be made to ensure that all rating/examining takes place under optimal conditions. Where possible, these conditions should be set, so that all examiners have an equal opportunity to perform at their best.
Moderation	This involves monitoring the performance of raters to ensure that they stay on level.
Analysis	Statistical analysis of all rater performances will ensure that individual candidates will not lose out in situations where examiners are either too harsh/lenient or are not behaving in a consistent manner. This is the part of Scoring Validity that is traditionally seen as reliability (i.e. the reliability of the scoring, or rating, system).
Raters	When we discuss the candidate (in terms of physical, psychological and experiential characteristics) we should also consider what we know of the examiners in terms of these same characteristics. Little research has been undertaken in which these have been systematically explored from the perspective of the rater.
Grading & Awarding	The systems that describe how the final grades are estimated and reported should be made as explicit as possible to ensure fairness. These are usually a combination of statistical analysis of results and qualitative analysis of the test itself.

Table 4 Scoring Validity

It is vitally important that the Criteria or Rating Scale we use in a test of writing or speaking should include criteria that reflect the model of language ability that we hypothesise reflects what exists in the mind of the test taker (for example the Cambridge ESOL rating scale should be *directly* related to the model of language ability shown above in Figure B3). This same model/set of criteria should also be reflected in the expected linguistic output of the test task. Without this triangulation we can never argue convincingly that our test is valid. I think of this relationship as the 'Golden Triangle' without which we can never claim that our test of speaking or writing is valid (see Figure 7).

Since it is a major decision to decide on the criteria that will be used for performance evaluation we will next focus on that aspect of development. The kind of scale (or rubric) to be used falls into one of two types, Holistic and Analytic. As with many other decisions that are made in language testing, the final decision as to which one to opt for is often down to practicality – for example it would be unwise to ask an examiner who is also the interlocutor to award scores on an Analytic scale since, as we will see below, it involves awarding multiple scores (so the person just may not have the time to get involved in such a complex task).

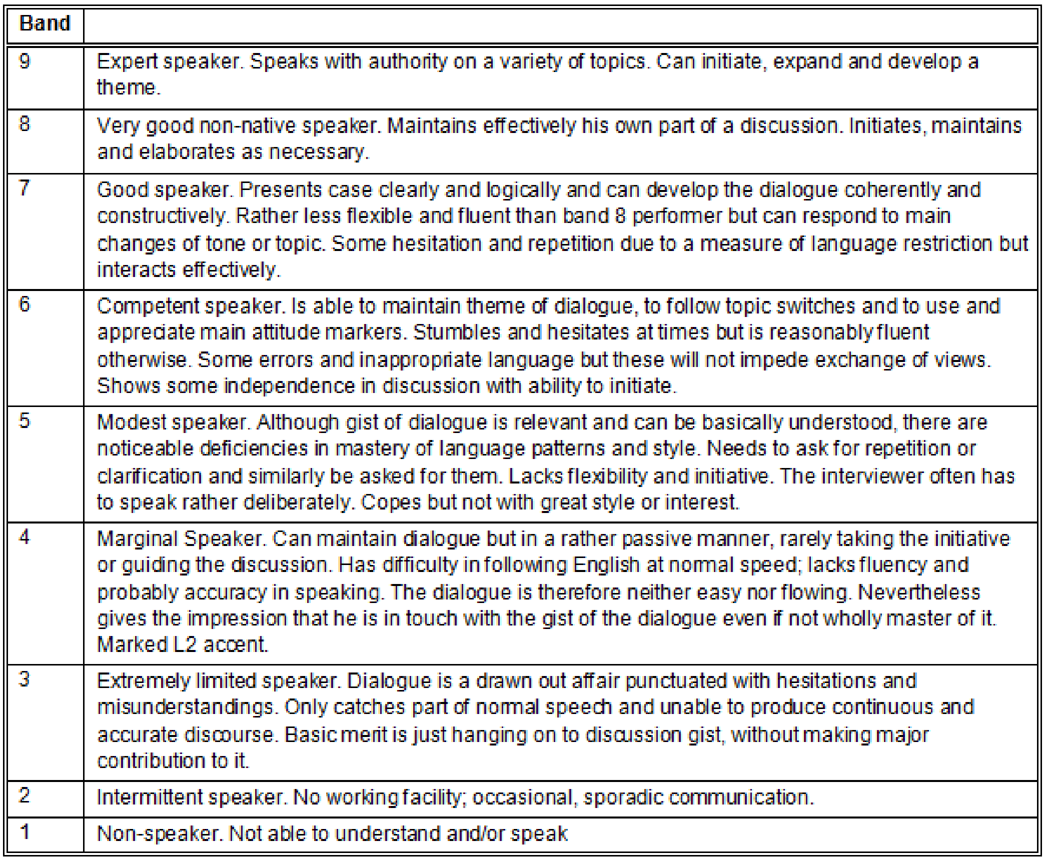
Figure 7 The 'Golden Triangle' relationship



5.1. The Holistic Rating Scale

In this type of scale the rater will award a single mark for the performance, based on a predetermined scale. An example of this type of scale is that of Carroll (1980) shown in Figure 7. The greatest advantage of the Holistic scale is its simplicity and speed. In addition, it is relatively easy to train raters to agree to within a band of the observed performance (this is the typical level of agreement set in standardisation procedures).

Figure 7   The Holistic Rating Scale (Carroll, 1980)



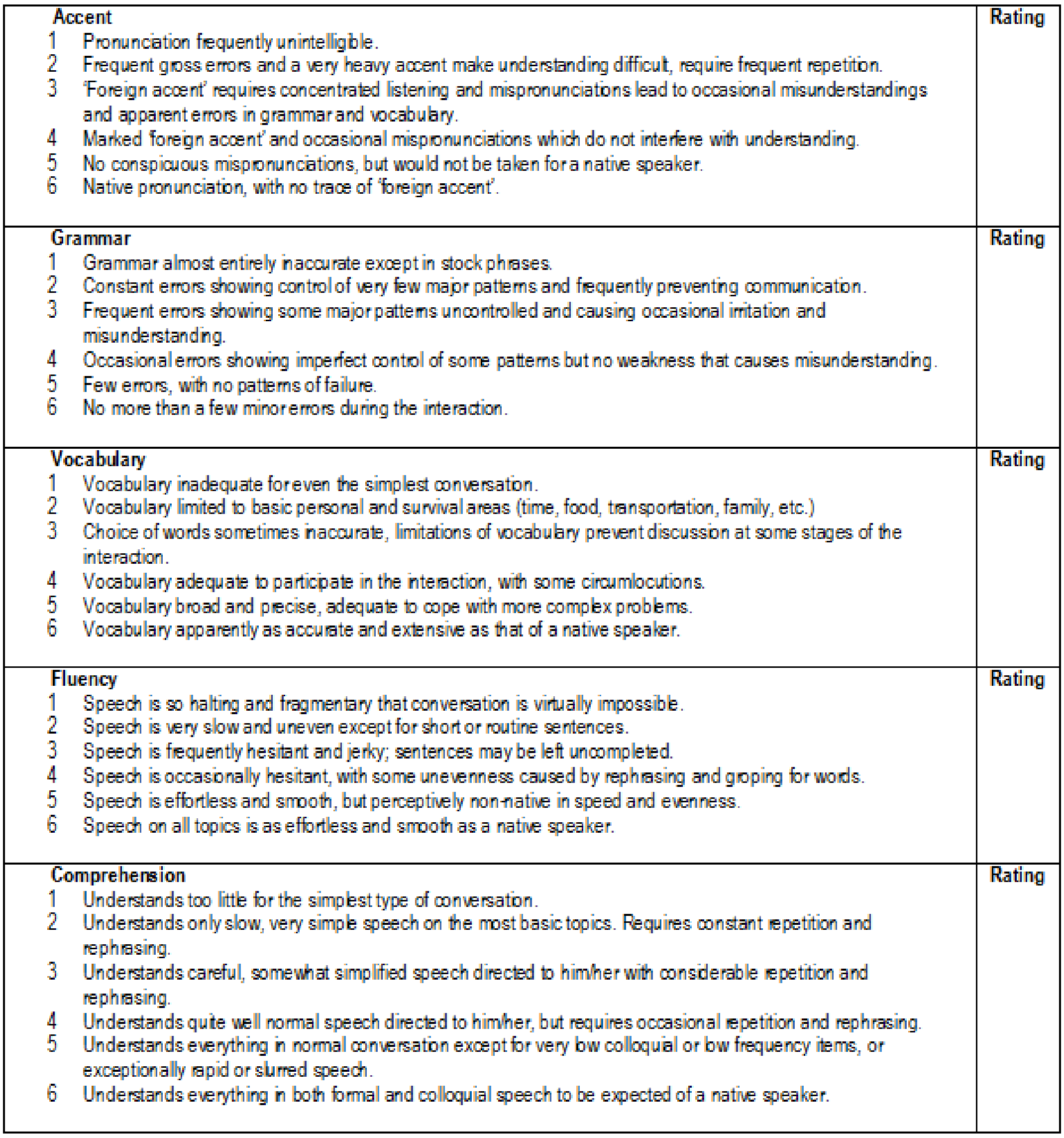
Band	
9	Expert speaker. Speaks with authority on a variety of topics. Can initiate, expand and develop a theme.
8	Very good non-native speaker. Maintains effectively his own part of a discussion. Initiates, maintains and elaborates as necessary.
7	Good speaker. Presents case clearly and logically and can develop the dialogue coherently and constructively. Rather less flexible and fluent than band 8 performer but can respond to main changes of tone or topic. Some hesitation and repetition due to a measure of language restriction but interacts effectively.
6	Competent speaker. Is able to maintain theme of dialogue, to follow topic switches and to use and appreciate main attitude markers. Stumbles and hesitates at times but is reasonably fluent otherwise. Some errors and inappropriate language but these will not impede exchange of views. Shows some independence in discussion with ability to initiate.
5	Modest speaker. Although gist of dialogue is relevant and can be basically understood, there are noticeable deficiencies in mastery of language patterns and style. Needs to ask for repetition or clarification and similarly be asked for them. Lacks flexibility and initiative. The interviewer often has to speak rather deliberately. Copes but not with great style or interest.
4	Marginal Speaker. Can maintain dialogue but in a rather passive manner, rarely taking the initiative or guiding the discussion. Has difficulty in following English at normal speed; lacks fluency and probably accuracy in speaking. The dialogue is therefore neither easy nor flowing. Nevertheless gives the impression that he is in touch with the gist of the dialogue even if not wholly master of it. Marked L2 accent.
3	Extremely limited speaker. Dialogue is a drawn out affair punctuated with hesitations and misunderstandings. Only catches part of normal speech and unable to produce continuous and accurate discourse. Basic merit is just hanging on to discussion gist, without making major contribution to it.
2	Intermittent speaker. No working facility; occasional, sporadic communication.
1	Non-speaker. Not able to understand and/or speak

However, the disadvantages of this scale include the danger of ‘trial by first impression’ meaning that since the examiner is asked to give one score only he or she may (and often does) simply rely on their first impression (or previous knowledge) of the candidate. So the score awarded may not actually reflect the observed performance. The other danger with this type of scale is that it represents at best a crude measure of the ability we are attempting to examine.

5.2. The Analytic Rating Scale

In this type of scale the developer first identifies the operations involved in responding to the task(s) and then attempts to create a marking scheme specifically to reflect these operations. This results in a multi-faceted scale, each component of which adds to an overall score. The most famous of all analytic scales is the Foreign Services Institute (FSI) scale, upon which most others have been based, see below.

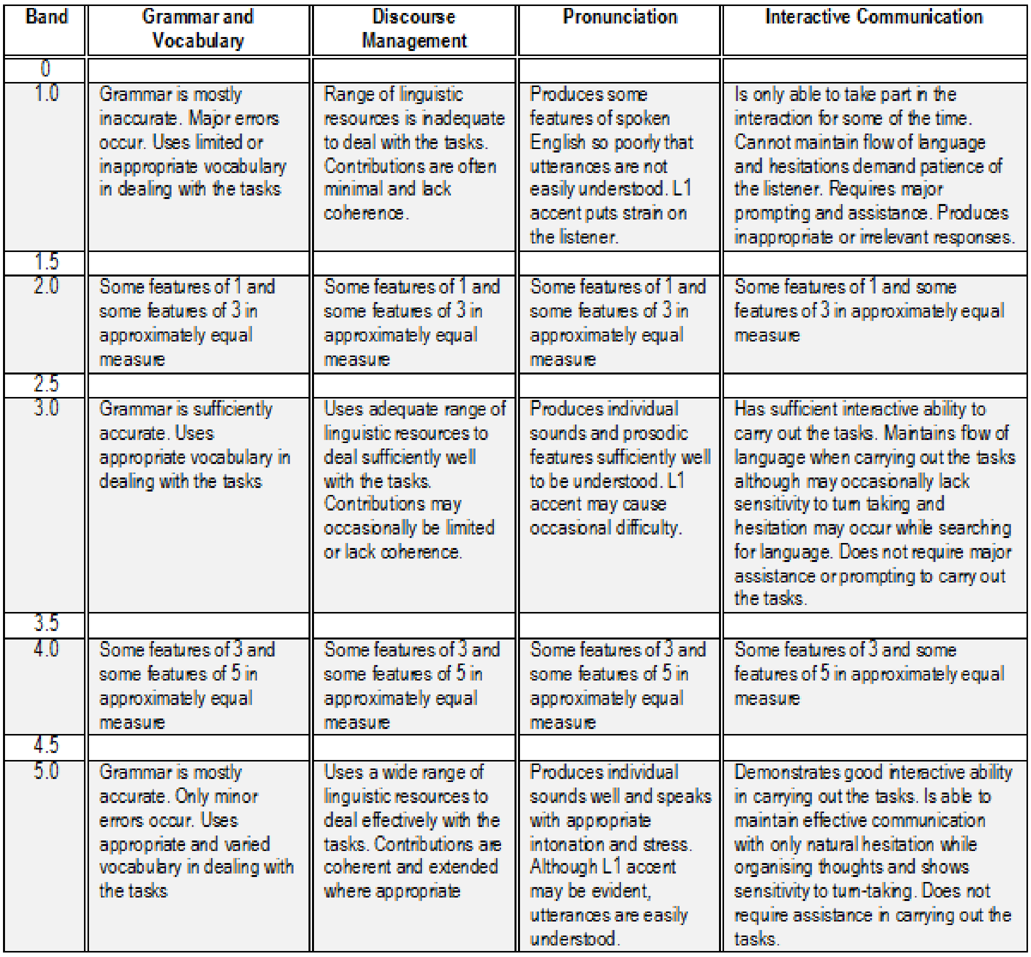
Figure 8 The Foreign Services Institute (FSI) Analytic Rating Scale



Accent 1 Pronunciation frequently unintelligible. 2 Frequent gross errors and a very heavy accent make understanding difficult, require frequent repetition. 3 'Foreign accent' requires concentrated listening and mispronunciations lead to occasional misunderstandings and apparent errors in grammar and vocabulary. 4 Marked 'foreign accent' and occasional mispronunciations which do not interfere with understanding. 5 No conspicuous mispronunciations, but would not be taken for a native speaker. 6 Native pronunciation, with no trace of 'foreign accent'.	**Rating**
Grammar 1 Grammar almost entirely inaccurate except in stock phrases. 2 Constant errors showing control of very few major patterns and frequently preventing communication. 3 Frequent errors showing some major patterns uncontrolled and causing occasional irritation and misunderstanding. 4 Occasional errors showing imperfect control of some patterns but no weakness that causes misunderstanding. 5 Few errors, with no patterns of failure. 6 No more than a few minor errors during the interaction.	**Rating**
Vocabulary 1 Vocabulary inadequate for even the simplest conversation. 2 Vocabulary limited to basic personal and survival areas (time, food, transportation, family, etc.) 3 Choice of words sometimes inaccurate, limitations of vocabulary prevent discussion at some stages of the interaction. 4 Vocabulary adequate to participate in the interaction, with some circumlocutions. 5 Vocabulary broad and precise, adequate to cope with more complex problems. 6 Vocabulary apparently as accurate and extensive as that of a native speaker.	**Rating**
Fluency 1 Speech is so halting and fragmentary that conversation is virtually impossible. 2 Speech is very slow and uneven except for short or routine sentences. 3 Speech is frequently hesitant and jerky; sentences may be left uncompleted. 4 Speech is occasionally hesitant, with some unevenness caused by rephrasing and groping for words. 5 Speech is effortless and smooth, but perceptively non-native in speed and evenness. 6 Speech on all topics is as effortless and smooth as a native speaker.	**Rating**
Comprehension 1 Understands too little for the simplest type of conversation. 2 Understands only slow, very simple speech on the most basic topics. Requires constant repetition and rephrasing. 3 Understands careful, somewhat simplified speech directed to him/her with considerable repetition and rephrasing. 4 Understands quite well normal speech directed to him/her, but requires occasional repetition and rephrasing. 5 Understands everything in normal conversation except for very low colloquial or low frequency items, or exceptionally rapid or slurred speech. 6 Understands everything in both formal and colloquial speech to be expected of a native speaker.	**Rating**

As can be seen in Figure 8, the Analytic scale is composed of a set of criteria. It is possible that these same criteria could be used to prepare an Holistic version of the scale, by collapsing the different criteria into a single band. The Analytic has been criticised for being simply a set of holistic scales – so that the distinction between the two is not at all as clear as we might first think. In this way, the Analytic scale suffers from the same disadvantages as the Holistic scale (magnified by the number of criteria included in the scale).

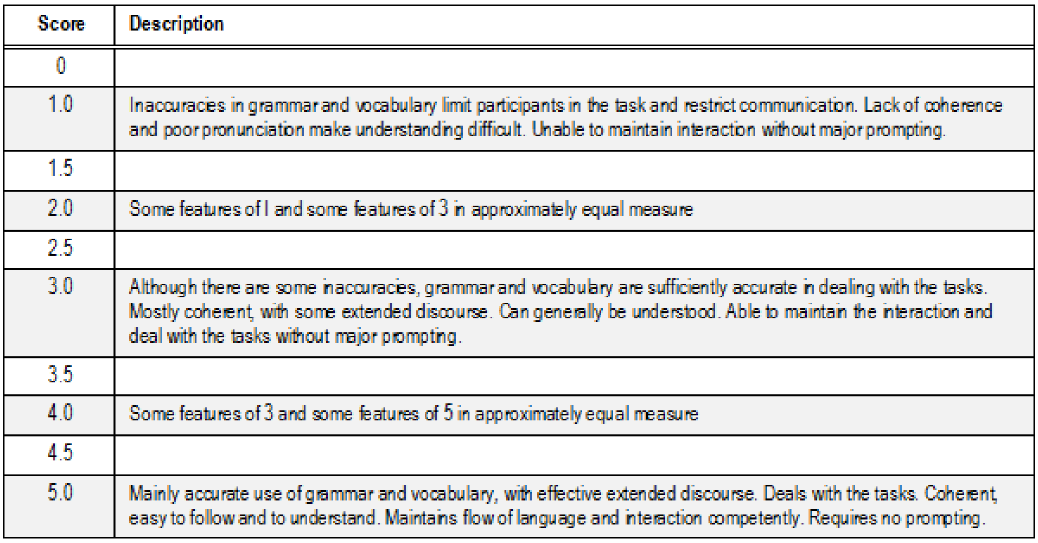
Figure 9 The Cambridge ESOL FCE Analytic Rating Scale



Band	Grammar and Vocabulary	Discourse Management	Pronunciation	Interactive Communication
0				
1.0	Grammar is mostly inaccurate. Major errors occur. Uses limited or inappropriate vocabulary in dealing with the tasks	Range of linguistic resources is inadequate to deal with the tasks. Contributions are often minimal and lack coherence.	Produces some features of spoken English so poorly that utterances are not easily understood. L1 accent puts strain on the listener.	Is only able to take part in the interaction for some of the time. Cannot maintain flow of language and hesitations demand patience of the listener. Requires major prompting and assistance. Produces inappropriate or irrelevant responses.
1.5				
2.0	Some features of 1 and some features of 3 in approximately equal measure	Some features of 1 and some features of 3 in approximately equal measure	Some features of 1 and some features of 3 in approximately equal measure	Some features of 1 and some features of 3 in approximately equal measure
2.5				
3.0	Grammar is sufficiently accurate. Uses appropriate vocabulary in dealing with the tasks	Uses adequate range of linguistic resources to deal sufficiently well with the tasks. Contributions may occasionally be limited or lack coherence.	Produces individual sounds and prosodic features sufficiently well to be understood. L1 accent may cause occasional difficulty.	Has sufficient interactive ability to carry out the tasks. Maintains flow of language when carrying out the tasks although may occasionally lack sensitivity to turn taking and hesitation may occur while searching for language. Does not require major assistance or prompting to carry out the tasks.
3.5				
4.0	Some features of 3 and some features of 5 in approximately equal measure	Some features of 3 and some features of 5 in approximately equal measure	Some features of 3 and some features of 5 in approximately equal measure	Some features of 3 and some features of 5 in approximately equal measure
4.5				
5.0	Grammar is mostly accurate. Only minor errors occur. Uses appropriate and varied vocabulary in dealing with the tasks	Uses a wide range of linguistic resources to deal effectively with the tasks. Contributions are coherent and extended where appropriate	Produces individual sounds well and speaks with appropriate intonation and stress. Although L1 accent may be evident, utterances are easily understood.	Demonstrates good interactive ability in carrying out the tasks. Is able to maintain effective communication with only natural hesitation while organising thoughts and shows sensitivity to turn-taking. Does not require assistance in carrying out the tasks.

In fact, this idea, of collapsing the basic elements of the analytic scale to make a single Holistic scale, been successfully attempted, as can be seen with the two FCE scales presented in Figures 9 and 10. As we saw in the overall description of the FCE test, these scales are designed to be used by the two examiners involved in each test event, the interlocutor using the Holistic (Figure 10) version and the Observer using the Analytic version (Figure 9).

Figure 10   The Cambridge ESOL FCE Holistic Rating Scale



Score	Description
0	
1.0	Inaccuracies in grammar and vocabulary limit participants in the task and restrict communication. Lack of coherence and poor pronunciation make understanding difficult. Unable to maintain interaction without major prompting.
1.5	
2.0	Some features of l and some features of 3 in approximately equal measure
2.5	
3.0	Although there are some inaccuracies, grammar and vocabulary are sufficiently accurate in dealing with the tasks. Mostly coherent, with some extended discourse. Can generally be understood. Able to maintain the interaction and deal with the tasks without major prompting.
3.5	
4.0	Some features of 3 and some features of 5 in approximately equal measure
4.5	
5.0	Mainly accurate use of grammar and vocabulary, with effective extended discourse. Deals with the tasks. Coherent, easy to follow and to understand. Maintains flow of language and interaction competently. Requires no prompting.

When reading the descriptors at levels 1, 3 or 5 we can see that there is an attempt to include all of the four criteria included in the analytic scale. However, we don’t really know if these criteria can be collapsed in this way — in that we don’t know for certain how the different criteria configure at each level. This fact represents the main advantage to using the Analytic scale, as it allows us to examine and score each of the criteria separately.

rater	candidate	Hol	GV	DM	Pr	IC	Tot
OM	1	4	**4**	**4**	**4**	**4**	20
OM	2	3.5	**3.5**	**3.5**	**3.5**	**3.5**	17.5
OM	3	4.5	4.5	4.5	5	4	22.50
OM	4	4	**4**	**4**	**4**	**4**	20
OM	5	3	**3**	**3**	**3**	**3**	15
OM	6	4	3.5	4	4	4	19.50
OM	7	2.5	2.5	2.5	3	3	13.50
OM	8	2.5	2.5	2.5	2	2.5	12
OW	1	4	**4**	**4**	**4**	**4**	20
OW	2	3	3	3	3.5	3	15.5
OW	3	2.5	2	2	2.5	2	11
OW	4	2.5	3	2.5	2.5	3	13.50
OW	5	3	**3**	**3**	**3**	**3**	15
OW	6	2.5	**2**	**2**	**2**	**2**	10.5
OW	7	4.5	**4.5**	**4.5**	**4.5**	**4.5**	22.5
OW	8	2	2	1.5	2	1.5	9
SJ	1	2.5	2.5	2.5	3	3	13.5
SJ	2	3.5	3	3	3.5	3.5	16.5
SJ	3	2.5	**2.5**	**2.5**	**2.5**	**2.5**	12.5
SJ	4	2.5	3	2.5	3.5	3	14.5
SJ	5	2.5	**2.5**	**2.5**	**2.5**	**2.5**	12.5
SJ	6	2.5	2	2	2.5	2.5	11.5
SJ	7	3.5	3.5	3	3	3.5	16.5
SJ	8	3	**3**	**3**	**3**	**3**	15

Table 5   Rating Data

However, there is a problem with this type of scale, as can be seen in the results shown below (Table 5) of an spoken test in which the FCE scale was used. Ignoring, for a moment the Holistic scores, it is clear from the table that on no fewer than 8 occasions (or 1/3 of the time) the rater awarded the same score for all four categories. This fact has two implications.

The first implication is that in one third of the cases the candidates have demonstrated the same level of ability on all four criteria — in other words the criteria for proficiency configure in the predicted way for all these subjects (a fact supported by the Holistic score matching the Analytic score exactly in all eleven cases). However, there is also a chance that the rater was simply assessing a single ability, and that he (in this particular case the examiners were men) filled in the same score in each category because of this. This is known as a *halo* effect.

Before finishing with this comparison of the two scale types, it is interesting to compare the results given (by two different 'independent' raters remember) using the two scales.

In the correlation matrix (Table 6), we would expect that the four elements of the Analytic scale would correlate highly with the total Analytic score and with the overall Total (after all they make up part of those scores). Similarly, we might expect that the Holistic score would correlate highly with the overall Total score. By correlation we mean the similarity in scoring patterns – all of the numbers in the table would be seen as being really very high.

What is interesting is the very high correlation between the Holistic score and the total Analytic score. This reflects the findings of a number of studies in which both analytic and holistic scores were given, and certainly suggests that both scales offer very similar outcomes.

	Holistic	**GV**	**DM**	**Pr**	**IC**	**Analytic**	**Total**
Hol	1.000						
GV	0.882	1.000					
DM	0.929	0.915	1.000				
Pr	0.869	0.881	0.891	1.000			
IC	0.883	0.865	0.904	0.828	1.000		
Analytic (GV+DM+Pr+IC)	0.934	0.958	0.973	0.942	0.944	1.000	
Tot	0.952	0.957	0.972	0.938	0.941	0.998	1.000

Table 6 Correlation Matrix

Reading

Nakamura, Yuji. 1996. Assessment of English Speaking Ability. *Journal of Humanities and Natural Sciences*, 102: 25-53.

An interesting overview of then current practice in Japan.

Kenyon, Dorry, M. and Tschirner, Erwin, T. 2000. The Rating of Direct and Semi-Direct Oral Proficiency Interviews: Comparing Performance at Lower Proficiency Levels. *The Modern Language Journal*, 88 (1): 85 – 101.

Thoughtful look at similarities and differences in how performances on the different test formats are scored.

O'Sullivan, Barry, Weir, Cyril J. and Saville Nick. 2002. Using observation checklists to validate speaking-test tasks. *Language Testing*, 19(1): 33-56.

Description of how a set of checklists used to monitor whether the functions predicted by the test developers were reflected in the language used by candidates when performing the tasks.

O'Sullivan, Barry. 2002. Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19(3): 277-295.

An experimental study into the effect on candidate performance in appeared speaking against of the level of acquaintanceship between the candidate and the interlocutor.

Further reading

Weir, C. 1993. *Understanding and Developing Language Tests*. Prentice Hall. Ch. 2

Hughes, A. 1989. *Testing for Language Teachers*. CUP. Ch. 10

Weir, C. 1988. *Communicative Language Testing*. Prentice Hall. Ch. 4 pp. 73-85

Weir, C. 2005. Language Testing and Validation: an evidence-based approach: Palgrave. Sections 7.3, 8.3 and 9.3.

References

BACHMAN, L.F. 1990. Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

BERRY, V. 2004. A study of the interaction between individual personality differences and oral performance test facets, Kings College, The University of London.

BORTFELD, H., LEON, S. D., BLOOM, J. E., SCHOBER, M. F. and BRENNAN, S. E. 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. Language and Speech, 44.123-47.

BROWN, A. 2003. Interviewer variation and the co-construction of speaking proficiency. Language Testing, 20.1-25.

BYGATE, M. 1987. Speaking. Oxford: Oxford University Press.

CARROLL, B. 1980. Testing communicative performance. Oxford: Pergamon.

DORNYEI, Z. and KORMOS, J. 1998. Problem-solving mechanisms in L2 communication: a psycholinguistic perspective. Studies in Second Language Acquisition, 20.349-85.

ELLIS, R. 2003. Task-based language learning and teaching. Oxford: Oxford University Press.

FODOR, J. A., BEVER, T. G., GARRETT, M. F. and . 1974. The Psychology of Language: An Introduction to Psycholinguistics and Generative Grammar. New York: McGraw-Hill.

FOSTER, P. and SKEHAN, P. 1996. The Influence of Planning and Task Type on Second Language Performance. Studies in Second Language Acquisition, 18.299-323.

—. 1999. The influence of source of planning and focus of planning on task-based performance. Language Teaching Research, 3.215-47.

FULCHER, GLENN. 1996. Does Thick Description Lead to Smart Tests? A Data-Based Approach to Rating Scale Construction. Language Testing, v13 n2 p208-38 Jul 1996.

HUGHES, R. 2002. Teaching and Researching Speaking. London: Longman.

KUNNAN, A. J. 1995. Test Taker Characteristics and Test Performance: A structural modelling approach.vol. 2: Studies in Language Testing. Cambridge: Cambridge ESOL & Cambridge University Press.

LEVELT, W. J. M. 1989. Speaking: from intention to articulation. Cambridge, MA: MIT Press.

LEVELT, W.J.M. 1999. Producing Spoken Language: a Blueprint of a Speaker. The Neurocognition of Language, ed. by C. M. Brown and P Hagoort, 83-112. Oxford: Oxford University Press.

MARSLEN-WILSON, W. and TYLER, L. K. . 1980. The Temporal Structure of Spoken Language Understanding. Cognition, 8.1-71.

MARSLEN-WILSON, W. D., TYLER, L. K. and SEIDENBERG, M. 1978. Sentence processing and the clause boundary. Studies in the perception of language, ed. by W. J. M. Levelt and F. d'Arcais. Hoboken NJ: John Wiley & Sons.

MCNAMARA, T. F. 1996. Measuring Second Language Performance. London: Longman.

MCNAMARA, T. F. and LUMLEY, TOM. 1997. The Effect of Interlocutor and Assessment Mode Variables in Overseas Assessments of Speaking Skills in Occupational Settings. Language Testing, 14.140-56.

NORRIS, J. M. , BROWN, J. D., HUDSON, T. and YOSHIOKA, J. 1998. Designing Second Language Performance Assessments. Technical Report #18. Hawai'i: University of Hawai'i Press.

NORTH, B. 1995. The Development of a Common Framework Scale of Descriptors of Language Proficiency Based on a Theory of Measurement. System, 23.445-65.

O'LOUGHLIN, K. 2002. The impact of gender in oral proficiency testing. Language Testing, 19.169-92.

O'LOUGHLIN, K. 2001. The equivalence of direct and semi-direct speaking tests. Cambridge: Cambridge University Press.

O'SULLIVAN, B., WEIR, C. J. and SAVILLE, N. 2002. Using Observation Checklists To Validate Speaking-Test Tasks. Language Testing, 19.33-56.

O'SULLIVAN, B. 1995. Oral Language Testing: Does the Age of the Interlocutor make a Difference?, Centre for Applied Language Studies, University of Reading.

—. 2000. Exploring gender and oral proficiency interview performance. System, 28.373-86.

—. 2002. Learner acquaintanceship and oral proficiency test pair-task performance. Language Testing, 19.277-95.

—. 2006. Modelling Performance in Oral Language Tests: Language Testing and Evaluation. Frankfurt: Peter Lang.

O'SULLIVAN, B. and PORTER, D. 1996. Speech Style, Gender and Oral Proficiency Interview Performance. RELC Conference. Singapore

O'SULLIVAN, B., WEIR, C. and HORAI, T. 2004. Exploring difficulty in speaking tasks: an intra-task perspective: Cambridge ESOL/The British Council/ IDA Australia: IELTS

ORTEGA, L. 1999. Planning and focus on form in L2 oral performance. Studies in Second Language Acquisition, 20.109-48.

PURPURA, J. E. 1999. Learner strategy use and performance on language tests: A structural equation modeling approach.vol. 8: Studies in Language Testing. Cambridge: Cambridge ESOL & Cambridge University Press.

TYLER, L. K. and MARSLEN-WILSON, W. 1977. The on-line effects of semantic context on syntactic processing. Journal of Verbal Learning and Verbal Behavior.683--92.

VAN LIER, L. 1989. Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. TESOL Quarterly, 23.489-508.

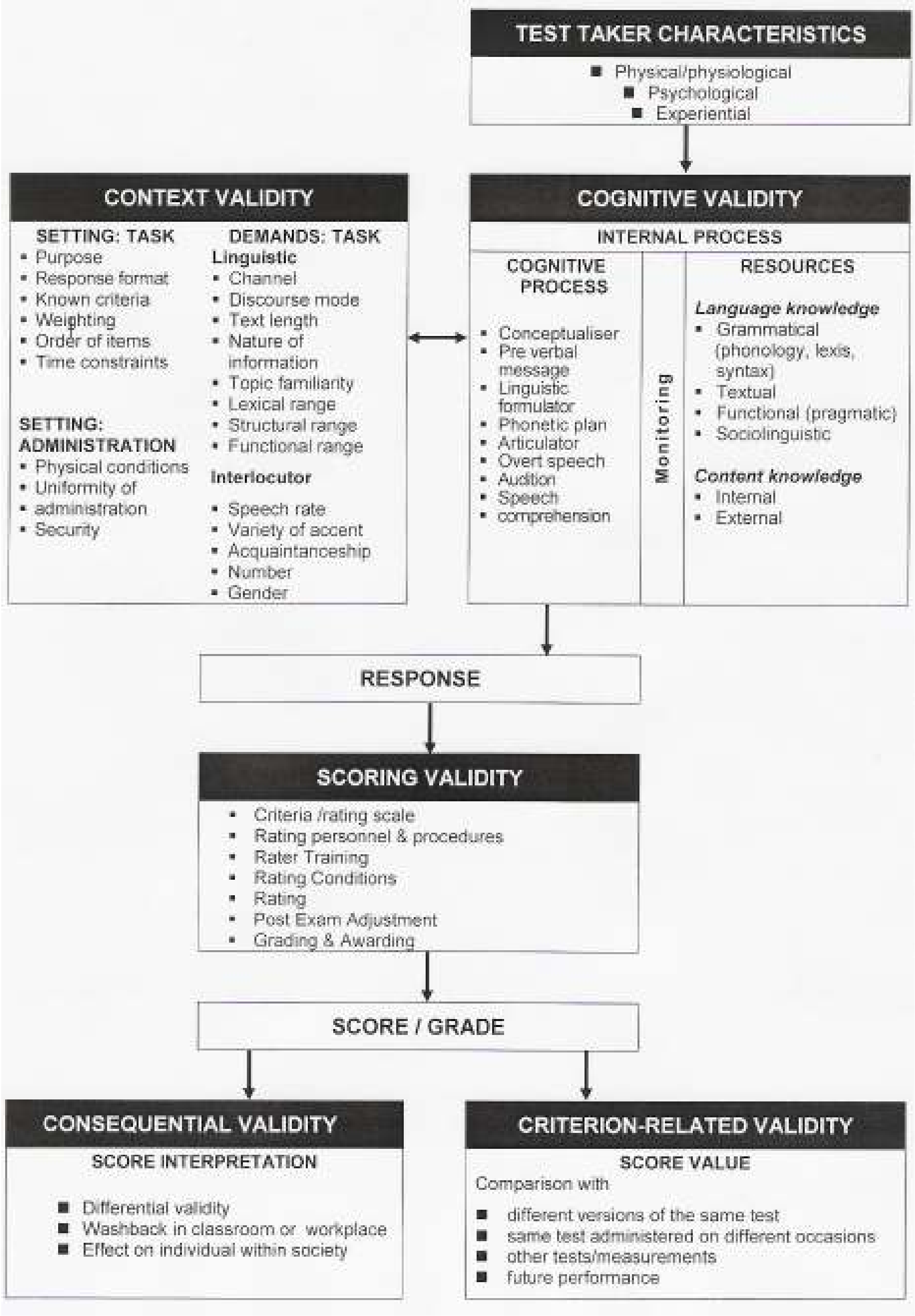
WEIR, C. J. 2005. Language Testing and Validation: an evidence-based approach. Oxford: Palgrave.

WEIR, C. J. 1993. Understanding and Developing Language Tests. Hemel Hempstead: Prentice Hall.

WIGGLESWORTH, G. 1993. Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. Language Testing, 10.305-36.

—. 1997. An investigation of planning time and proficiency level on oral test discourse. Language Testing, 14.85-106.

Appendix 1 Validation Framework for Speaking



Appendix 2 Guidelines for Speaking Tests

	Instructions
1	Are the instructions clear on what students have to do?
2	Are the instructions written at a level clearly below that expected of the candidates?
3	Are the instructions grammatically correct?
4	Are the instructions spelled correctly?
5	Are the instructions likely to be familiar to the students?
6	Are the instructions specific about the amount of planning time allowed for each task?
7	Are the instructions specific about the amount of speaking time allowed for each task?
8	Do students know the assessment criteria (rubric)?
	WRITING TASKS
1	Does the task measure what it is supposed to measure? Make sure task types are suitable for testing the specified functions.
2	Do the tasks appropriately sample the range of speaking ability expected at this level?
3	Is each task closely related to real-life language use? Try to make it as realistic as possible.
4	Are visual stimuli, e.g. pictures, drawings, tabled data, etc., clear and accessible? Does the test avoid visual and mental overload?
5	Are the tasks at the right level of difficulty? a. Is the type of drawings/ pictures/ information familiar to the students?
	b. Are the tasks familiar to the students? Have the students likely to have practised the same type of tasks?
	c. Are the topics sufficiently familiar so every student has enough knowledge to write about? Topics should not be biased in any way.
	d. Is the length of output appropriate to the stage? The length of speaking required should not be too much for the student.
	e. Is time given sufficient to understand the question and deliver a satisfactory response? Danger with giving too much or too little time.
	f. Does the test include a variety of questions for both good and weak students? They are necessary for making differentiation between students. Simple or easier tasks/items should be given first and more difficult tasks later.
6	Is there a choice of task? If so, are you sure they are equivalent in all respects? Normally it is better not to give a choice to be fair to the students.
	RATING SCALE (RUBRIC)
1	Do the criteria contained in the scale match the expectations of the task designer? If the task is designed to measure one aspect of language this must be reflected in the scale.
2	Are the descriptors written in clear and unambiguous language?
3	Is it easy to compute marks to generate the final score? Ideally, raters should not be asked to perform any calculations.

4	4. What is the pass score? How do you decide on this?
5	5. Are marking and markers reliable? a. Have you ensured that all raters fully understand the scale?
	b. Are all the markers aware of and agreed on critical boundary?
	c. Are all markers standardized to these criteria? They should be. It is useful to have samples of speaking at different levels to illustrate different performances in respect of each of the criteria. This will help with reliability of marking between teachers as well as for the individual teacher over time.
	d. Is the marker consistent in his/her own standard of marking?
	FINAL PRE-IMPLEMENTATION CHECKS
1	Is it clear to the students what the individual parts of the test are testing? Are they told what each task tests?
2	Have you proof-read the test? Be sure to eliminate any mistakes by reading over the final version at least twice. The more times you read it, the better. Check that any visual input has been prepared to a high level of quality.
3	Have you given or are you going to give tests and marking schemes to interested, trustworthy, professional colleagues for their comments? You should! Test of speaking should not be a solitary activity.
4	Have you checked that the kind of language (i.e. functions) predicted at the development phase actually occur in the operational phase!