7 Multiplication and division
Trends in language assessment

Introduction

Personal reflection

How do you think language assessment may have changed over the past 100 years?
What do you think might have changed in relation to ....
Who takes language assessments and why?
What techniques are used to assess language abilities?

A little over 50 years ago, in the spring of 1961, a conference was held in Washington, DC, to discuss options for a new test of English for international students coming to study at American universities. As the delegates settled back in their chairs, John Carroll, a leading psychologist and linguist, began his opening address with the words: ‘Language testing has a long history, and much experience and wisdom have accumulated ... [it] has reached such a stage of professionalization that a whole book on the subject is about to appear’ (Carroll, 1961, p. 313). Today, when we can point to the shelves of books on language assessment that appear each year as well as to university courses,
The development of the profession would take time. Following publication of Lado's ground-breaking *Language Testing* in 1961, the teaching of the subject as a branch of applied linguistics was pioneered over the course of the 1960s by Alan Davies and Elisabeth Ingram at the University of Edinburgh. It was not until 1984, however, that the first academic journal dedicated to language assessment, also titled *Language Testing*, was launched. The International Language Testing Association (ILTA), the global association for language testing and assessment practitioners and scholars, was set up in 1992.

The use of the word 'testing' rather than 'assessment' is worth noting because it reflects the early focus of attention. It was only really from the 1990s that research and guidance concerning classroom-based language assessment by teachers and learners would begin to be published in any quantity in international journals. In spite of the shifts in perspective over the intervening half century, the issues raised by Lado and Carroll remain relevant. Many of the same questions that exercised language testers in 1961 are still widely discussed today.

The issues that have attracted the most attention over the intervening years have generally been technical questions concerning what it is that is being assessed (i.e., what is 'language proficiency' or 'language ability') and how better to assess it. Davies (2008a), describing the changes in textbooks used to teach about language testing and assessment since those early days, called these matters of 'skills and knowledge'. Skills include appropriate testing and analysis techniques: item writing, test delivery systems and statistical test analysis. Knowledge covers theories for language description and measurement and may involve exploring different models of language learning and approaches to language teaching, as well as language assessment.

More recently, scholars have become increasingly interested in questions of values: the who and why of assessment. Davies called these 'principles'. He placed in this category issues around the use of language tests, their fairness and impact. These involve the responsibilities of language assessors, the ethical choices they make and the impact of their work upon society. Who are the people or agencies that use language assessments? For what purposes do they use these assessments? What uses of assessment can be supported on ethical grounds? What effects do the use and misuse of assessment have on educational systems? What other social consequences result when tests and assessments are introduced?

Chapter 4 introduced the idea that all assessments involve constructs and showed the need for assessment developers to define, describe and justify the knowledge, skills or abilities they intend to assess. Like other constructs that psychologists have tried to measure (such as intelligence, critical thinking or social development), there has been a great deal of disagreement about the existence and nature of language ability, and views have shifted over time. Developers of language assessments generally look for established theories to guide them in deciding what to target in their tests. But different theoretical accounts of language and different theories of measurement have come in and out of favour in different parts of the world. In trying to summarise the trends, Spolsky (1977) described four overlapping phases in language testing (1 to 4 in Table 7.1) and these can be mapped more or less directly to trends or tendencies in language teaching. I have added two further categories (5 and 6) because they are more closely associated with classroom-based assessment by teachers, while Spolsky was concerned with more formal testing, generally at the institutional level or above.

Since the 1980s, as more attention has been given to the history of language assessment, it has become clearer that it is rarely possible to fit operational assessments neatly into any one of these categories. On the other hand, Spolsky's distinctions do point towards some of the key influences and broad historical developments within a more complex reality. Language assessment has been shaped by a wide range of influences, including practicality, political expediency and established customs, as well as developments in language teaching, applied linguistics and other allied disciplines such as educational psychology. Global trends including the growth in international trade, mass migration and tourism have brought new reasons for learning and using languages, and naturally assessment has also been affected by these broader social changes.
Table 7.1 Phases (or tendencies) in language assessment

<table>
<thead>
<tr>
<th>Language testing</th>
<th>Language teaching</th>
<th>Favoured techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Pre-scientific/</td>
<td>Grammar translation</td>
<td>Translation, grammar exercises, essays.</td>
</tr>
<tr>
<td>traditional</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Psychometric-</td>
<td>Audio-lingualism</td>
<td>Multiple-choice tests of grammar, vocabulary, phonetic discrimination, reading and listening comprehension. Focus on reliability.</td>
</tr>
<tr>
<td>structuralist</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Psycholinguistic</td>
<td>Natural approach</td>
<td>Cloze, dictation.</td>
</tr>
<tr>
<td>socio-linguistic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Communicative</td>
<td>Communicative/</td>
<td>Assessment tasks intended to reflect 'real life' language use. Integrated skills. Focus on validity of content.</td>
</tr>
<tr>
<td>task based approach</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Formative testing</td>
<td>Behavioural objective and mastery learning approaches</td>
<td>Tests that are sensitive to instruction and inform remediation. Focus on the products of learning.</td>
</tr>
</tbody>
</table>

Although Spolsky described tests such as the Certificate of Proficiency in English (CPE) (University of Cambridge Local Examinations Syndicate 1913) as 'traditional' and 'pre-scientific' as though the practices it embodied dated from time immemorial, the use of formal language tests was a relatively recent innovation. The test represented the current state-of-the-art linguistic thought of the day. Formal written tests had been in operation in China for centuries (Miyazaki, 1981), but they had only been in widespread use in Europe and America for a few decades. The established tradition there was for teachers to examine students orally on what they had been taught. This tradition still persists in the *viva voce* examination of doctoral candidates around the world and was widely used until very recently in Russia and other countries in the Soviet sphere of influence for school examinations. Where significant rates of failure did not suit the interests of teachers or their superiors, assessments of this kind often came to serve a largely ceremonial purpose: endowing the completion of a course of study rather than providing a rigorous test of knowledge, skills or abilities.

It is interesting to note that although much has altered in the way we have thought about language, learning and assessment since the beginning of the twentieth century and a variety of techniques have come in and out of favour, the majority of the test formats that were used in the CPE (most of which were already well established in tests of modern and classical languages) are still in use in operational tests today (see Table 7.2). Essays, reading aloud and dictation appear in the Pearson Test of English – Academic (PTE-A) introduced by Pearson Language Assessments in 2010.

A practical handbook in language assessment for teachers written in the late 1960s reflected the three objectives of traditional language programmes in formal schooling that predominated for most of the last century: 'to enjoy the literature written in the target language, to appreciate the culture of the target country and especially to converse freely with its people' (Valette, 1967, p. 4).

There was, perhaps inevitably, a certain tension between these goals. The older ideal of language education was for learners to develop an appreciation of the finest in the foreign literature and culture. This aim has gradually been displaced in many Ministries of Education and other policy-making bodies by the more utilitarian view that knowledge of foreign languages is a basic skill that has economic value: readying workers to participate in international markets. Learners themselves often wish to acquire a language not so much to access the cultural highlights as to help them to travel, to build a new life in a foreign country, to access technical information, to do business or to participate in a religion.

Today, in many parts of the world, literature has lost its former status as a natural component of any language programme and is not so often found in language tests. It does retain a foothold. A literature component

Table 7.2 The format of the 1913 version of the Certificate of Proficiency in English

<table>
<thead>
<tr>
<th>Paper</th>
<th>Time (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written</td>
<td></td>
</tr>
<tr>
<td>Translation into French or German</td>
<td>2.0 hrs</td>
</tr>
<tr>
<td>Translation into English and grammar</td>
<td>2.5 hrs</td>
</tr>
<tr>
<td>English essay</td>
<td>2.0 hrs</td>
</tr>
<tr>
<td>English literature</td>
<td>3.0 hrs</td>
</tr>
<tr>
<td>English phonetics</td>
<td>1.5 hrs</td>
</tr>
<tr>
<td>Oral</td>
<td></td>
</tr>
<tr>
<td>Dictation</td>
<td>0.5 hrs</td>
</tr>
<tr>
<td>Reading aloud and conversation</td>
<td>0.5 hrs</td>
</tr>
</tbody>
</table>

can still be found today in the CPE as one part of the current writing paper. However, as early as the 1930s, the University of Cambridge came under pressure from CPE test users to address business-oriented uses of language. In response, a paper in 'economic and commercial knowledge' was offered from 1936 as an alternative to literature (Weir, 2003).

It is revealing to see which formats from the modern language assessment toolkit were not included on the 1913 paper. There were no multiple choice questions and no tests of reading or listening comprehension. Multiple choice questions were absent partly for the very good reason that they had only just been invented -- they did not come into widespread use in language tests until the 1920s and then mostly in the US. The grammar section of the 1913 CPE involved correcting sentence errors and composing sentences to demonstrate the meaning of given vocabulary items. Reading skills were addressed through translation from English and listening skills through dictation. It would take another 50 years for papers titled 'Reading Comprehension' and 'Listening Comprehension' (with multiple choice questions) to appear in later versions of the CPE.

**Task 7.3**

Look at the sample items from the 1913 CPE in Example 7.1. How is it different from any more recent tests you are familiar with? What criticisms would you make of the test?

**Example 7.1 Extracts from the Certificate of Proficiency in English (CPE) (June 1913)**

**Translation into French**

My grandfather's recollections of Culloden were merely those of an observant boy of fourteen, who had witnessed the battle from a distance. The day, he has told me, was drizzly and thick: and on reaching the brow of the Hill of Cromarty, where he found many of his townsfolk already assembled, he could scarce see the opposite land. But the fog gradually cleared away; first one hill-top came into view and then another: till at length the long range of coast, from the opening of the great Caledonian valley to the promontory of Burghead, was dimly visible through the haze ....

**Grammar**

1. Give the past tense and past participle of each of the following verbs, dividing them into strong and weak; add explanations: tell, wake, buy, eat, lay, lie ....

---

5. Correct or justify four of the following sentences, giving your reasons:
   (a) I hope you are determined to seriously improve.
   (b) Comparing Shakespeare with Aeschylus, the former is by no means inferior to the latter.

---

**English phonetics**

1. Describe fully the articulation of the various vowel sounds in the (ordinary) spelling of which the letter o is used (alone or in combination) in the above passage.

---

**English essay**

(Two hours)

Write an essay on one of the following subjects:
(a) the effect of political movements upon nineteenth century literature in England;
(b) English Pre-Raphaelitism;
(c) Elizabethan travel and discovery;
(d) the Indian mutiny;
(e) the development of local self-government;
(f) Matthew Arnold.

---

Every established testing system has its detractors and Strevens (1965, p. 7) listed three sources of frustration with the CPE and other similar examinations. These included that they are 'poor in validity and reliability as devices for the assessment of performance; that they incorporate a non-authentic and stilted form of English; and above all, that they perpetuate styles of teaching which are educationally undesirable ... preparation for these examinations takes the place of learning to use English in an effective way'.

In spite of its name (Certificate of Proficiency in English), the CPE was not a proficiency test in the sense used in this book, but was always intended as a form of achievement test. It was based quite closely on a model of language teaching; Weir (2003) notes that the content of the 1913 test closely mirrors a popular language course book of the time (Sweet's 1899 *The Practical Study of Languages*). There was very little difference between the exercises found in text books and the tasks presented in tests.
By the 1960s, this model was seen by reformers such as Strevens to be severely outdated. The CPE was encouraging translation and essay writing in the classroom where Strevens wanted to promote audio-lingual methods involving greater use of spoken language. The reformers argued that there needed to be a radical departure from the established approach to learning languages as an academic school exercise towards a more practical concern with ‘basic communication with native speakers’ (Lado, 1961, p. 1). The cure for this negative and conservative washback effect was, Strevens suggested, to cut the connection between test and curriculum, freeing teachers to employ the new methods without worrying about what would appear on the test. This, it was argued, could be done by having test tasks that were very unlike teaching tasks. The poor standards of validity and reliability should be addressed by building stronger links with linguistic theory and introducing item types that could be scored in a consistent way by untrained markers.

The alternative that appealed to Strevens was still called at that time the 'new type' of objective examination. In language assessment, this implied what Spolsky (1977) called the psychometric-structuralist phase. Its influence on language assessment in the US had grown through the 1930s and 1940s. By the 1960s it was gradually being taken up in many parts of the world, sometimes in the face of strong resistance from traditionalists.

The approach is well exemplified in Lado's (1961) book. The structuralist element came from the predominant linguistic theory in the US at the time associated with the standard textbook, Bloomfield's *Language* (1933). Lado brought together Bloomfield's 'linguistic science' with the technology of psychometric testing to build what he claimed was a more scientific way of assessing language. The psychometric element involved a concern for the technical measurement qualities of tests. Arguments for the validity of traditional tests were based on the qualifications and standing of the people who designed and built them. In contrast, psychometric approaches to validity emphasised the need for hard statistical evidence that a test was working effectively. Such evidence might include high levels of reliability and similarities in the results obtained from different tests of the same abilities (concurrent validity). Objective scoring was important and so the use of multiple choice question formats was favoured over essays because they eliminated subjective judgements about whether an answer was right or wrong.

Important papers by Cronbach and Meehl (1955) and Campbell and Fiske (1959) had recently argued for a more rigorous approach to validity that took account of test method. This was the multi-trait, multi-method approach to validation. Campbell and Fiske suggested that tests of the same ability employing different methods should give similar results (convergent validity), but that tests of different abilities that employed the same method should give different results (divergent validity). If this was found to be the case, it would provide strong evidence for validity. Let’s say, for example, that we have two reliable tests of grammatical knowledge: one using multiple choice questions and the other using a very different format, such as samples of writing. If they give very similar results, but a test of grammar and a test of vocabulary that both use multiple choice questions give very different results, this would strongly suggest, first, that both ‘grammar’ and ‘vocabulary’ were distinct abilities and, second, that the tests under study were effective ways of measuring these abilities. If, on the other hand, the results of the multiple choice tests were similar to each other, but very different to the writing samples, this would suggest that a method effect was playing a greater part in performance than the intended abilities and that the test designers should reconsider their approach.

In Lado's linguistic model, sometimes known as a skills/components model, language was made up of linguistic components and the 'total skills' of listening, speaking, reading and writing (see Table 7.3). The linguistic components included pronunciation (sound segments, intonation and stress), grammatical structure (syntax and morphology), the lexicon and another element not as widely included in psychometric language tests and so excluded from Carroll’s similar schematic in Table 7.3: ‘cultural meanings’.

In what is sometimes called discrete-point testing, Lado recommended that each of the components should be tested separately. However, he also acknowledged that components would always occur in combination in language use, which involves the ‘total’ skills of listening, speaking, reading and writing and that tests of these might also have value in some circumstances. Valette’s (1967, p. 18) recommendations on preparing texts for use in a classroom test of listening comprehension show how a discrete-point approach to

<table>
<thead>
<tr>
<th>Skill</th>
<th>Language aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditory comprehension</td>
<td>Phonology or orthography</td>
</tr>
<tr>
<td>Oral production</td>
<td>Morphology</td>
</tr>
<tr>
<td>Reading</td>
<td>Syntax</td>
</tr>
<tr>
<td>Writing</td>
<td>Lexicon</td>
</tr>
</tbody>
</table>

Table 7.3 Grid for measuring language abilities from Carroll (1961)
language affected ‘total skills’ testing: ‘the teacher should, a couple of
days in advance, prepare a first draft of the script ... making sure that
he has used all the structures and vocabulary used that week’; the
priority was not so much to capture language in use in realistic
situations, but to exemplify key linguistic components.

The linguist Bloomfield (1933), influenced by behaviourist
psychology, believed that learning a language was largely a matter of
imprinting sounds, words and sentence patterns in the mind through
habit formation: ‘the thousands of morphemes and tagmemes of the
foreign language can be mastered only by constant repetition ... Every
form that is introduced should be repeated many times’ (Bloomfield,
1933, p. 506). In classroom assessment, the emphasis was on accuracy.
Accurate language should be rewarded with enthusiastic praise
because this was believed to reinforce good habits. Incorrect forms
should be quickly put right and the learner asked to imitate and repeat
the correct form. The hardest habits to imprint would be those that
involved forms that were very different than those found in the
learner’s native language.

Lado argued that discrete-point tests focusing on problems of this
kind would be the most efficient, but acknowledged the practical need
for tests that could work effectively for learners from different first
language groups. Constructing a comprehensive test of all elements of
a language would be entirely unrealistic, but tests of skills that involved
a more or less random sample of a few sounds or structures should be
effective in giving a more general picture of proficiency. Such tests
might work reasonably well, he suggested, to indicate whether or not
a learner had enough ability in a language to allow them to use it in
studying other subjects.

In traditional tests of grammar, vocabulary or phonetics such as the
1913 CPE, assesses stated grammar rules, gave definitions and
produced phonetic transcriptions. Lado rejected these techniques on
the grounds that knowing rules would not necessarily equate to being
able to apply them in real world communication. Instead, he insisted
that test takers should demonstrate the ability to apply what they
knew to real samples of language. Lado also disapproved of other
traditional methods used in the 1913 CPE – translation, essays and
dictations – for a variety of reasons, both theoretical and practical.
With regard to translation, Bloomfield (1933, p. 506) had observed
that ‘the meaning of foreign forms is hard to convey. Translation into
the foreign language is bound to mislead the learner, because the
semantic units of different languages do not match.’ He pointed out
that the scoring of translations was problematic as different raters may
have different opinions about what constitutes a good rendition.
Accurate word-for-word translation may misrepresent the overall
meaning of a text while a reasonable rendition of the gist may be faulty
at the sentence level. Translation was fine for tests of translation skills,
but not in tests of general language proficiency.

Lado argued that striking the appropriate balance between tests of
components and tests of skills was a matter best decided on a case-by-
case basis according to the purpose of the assessment. However, he
believed that discrete-point testing should generally be given priority
because it would be more efficient and more comprehensive. Language
points for inclusion could be selected on a rational basis and a relatively
wider range could be dealt with in less time than would be possible
through traditional techniques. Essays, for example, would not
generally require the use of question or request forms while dictations
could not test awareness of word order or vocabulary (because both
are provided by the speaker who reads the input text). He also argued
against the use of tests of reading and listening comprehension based
on input texts taken from real life settings on the grounds that they
would introduce ‘extraneous factors’ (Lado, 1961, p. 205) that were
not strictly linguistic, such as technical knowledge, literary appreciation
or general intelligence.

In contrast to these traditional methods, Lado favoured tests made
up of collections of large numbers of very short, unrelated items. This
would permit the careful test developer to assess a relatively wide
range of components of language in a comparatively short time. He
particularly favoured basing both teaching and assessment on
contrastive analysis – the identification of differences between a
learner’s native language and the language being learned. These
differences could be used to predict the problems that learners would
have when taking up a new language. For example, Spanish speakers
learning English find it difficult to distinguish between the ‘d’ sound in
‘day’ and the ‘th’ sound in ‘they’ (phonetic /d/ and /θ/), because there is
no such contrast in Spanish. For similar reasons, Japanese speakers
have trouble dealing with post-modifiers and relative clauses in English
(sometimes preferring incorrect patterns – ‘the doesn’t work clock’ –
to the relative clause – ‘the clock which doesn’t work’).

Lado (1961, p. 205) argued that ‘it is more economical and will
result in more complete sampling to work from the language problems
and then to seek situations in which particular problems can be tested’.
Multiple choice questions were to be preferred because scoring was
objective, requiring no judgement on the part of the scorer, who
would simply match the response to the answer key provided. Such
tests could be very quickly and reliably scored by untrained clerical
staff or by the newly available electrographic machines. The first
commercially available Test Scoring Machine, invented by a high
school physics teacher, was introduced as the IBM 805 in 1936 and
had demonstrated its value in the large-scale testing of army recruits
during World War II (Fulcher, 2010).
Example 7.2 Discrete point items based on Lado (1961)

Skill – Auditory comprehension: Language aspect – Phonology: Sound segments

The examiner reads sets of very similar sentences.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A Sam got the fool back.</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>B Sam got the fool back.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C Sam got the full back.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>A We drove the sheep.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

On the answer sheet, the test taker places a mark against the sentences which sound the same in each set:

1 A&B [ ] 2 A&B [ ] 3 A&B [X]
1 A&C [ ] 2 A&C [X] 3 A&C [ ]
1 B&C [ ] 2 B&C [ ] 3 B&C [ ]
1 A,B&C [ ] 2 A,B&C [ ] 3 A,B&C [ ]
None [X] None [ ] None [ ]

Skill – Reading: Language aspect – Syntax: Head of a subject construction (Lado, 1961, p. 158)

The friend of the teacher that John met when he visited his mother and his aunt visited the library today.

A person visited the library. This person was:
1 the friend
2 the teacher
3 John
4 the mother
5 the aunt

Skill – Writing: Language aspect – Syntax: Modification

Complete the sentence by rearranging the choices.

'__________ visited us today.'

1 Teacher's
2 That John met
3 The
4 When he went home
5 Friend

Skill: Auditory comprehension/(alternatively reading comprehension) (Lado, 1961, p. 213)

The test taker hears (alternatively reads): *Have you seen our new house dog?*

Figure 7.1 Auditory comprehension


The influence of the skills/components approach can be seen in the test of English for international students that eventually emerged from the process set in train by the 1961 conference: the *Test of English as a Foreign Language (TOEFL®)*. Introduced in 1964, the TOEFL, alongside more discrete-point tests of 'structure and written expression' and vocabulary, included tests of reading comprehension and listening (see Table 7.4). All parts of the test were scored objectively: the subjectively scored elements of the traditional language examination – essays and tests of speaking – were excluded.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Time limit (minutes)</th>
<th>Number of questions</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening</td>
<td>30-40 minutes</td>
<td>50</td>
<td>20 statements with paraphrasing response options, 15 brief dialogues with comprehension questions, 15 questions based on 5 minute lecture</td>
</tr>
<tr>
<td>Structure and written expression</td>
<td>25 minutes</td>
<td>40</td>
<td>15 sentence correction items, 25 paragraph completion items</td>
</tr>
<tr>
<td>Reading comprehension</td>
<td>55 minutes</td>
<td>50</td>
<td>5 tests (100-250 words each) from social science, fiction and natural science texts: 4/5 items with each testing: (a) specific facts; (b) interpretation of facts and sentences; (c) understanding the passage (inference and tone).</td>
</tr>
<tr>
<td>Vocabulary</td>
<td></td>
<td></td>
<td>Vocabulary: sentence completion and phrase synonyms</td>
</tr>
</tbody>
</table>

Source: based on Spolsky (1995)
The psycholinguistic-socio-linguistic phase

After 1964, the TOEFL expanded and thrived while employing much the same format for the next 20 years. Nevertheless, like CPE, it attracted its share of criticism from teachers. Traynor (1985) voiced the concern that the absence of speaking and writing tasks in the test would lead to a neglect of these skills in teaching. In response to such criticism, Writing (essay) and Speaking tests were first offered as optional extras before being fully incorporated into the test in 2005. Traynor was also troubled that many items seemed to require knowledge of North American culture. Johnson et al. (2007, p. 75) reported the apprehension of the teacher in their study that multiple choice items of the kind used in the TOEFL 'trick' the unwary test taker into giving incorrect responses and (echoing Stevens's criticism of the CPE) that the language of the test 'does not always correspond to everyday language use'.

During the 1960s and 1970s, applied linguists reacted against both Bloomfield's structuralist linguistics and behaviourist learning models of the kind favoured by Lado. Evidence emerged that contrastive analysis - identifying key differences between language systems - was inadequate as a means of predicting learning problems. Learners often made errors that could not be traced to differences between languages and, in fact, learners from very different language backgrounds seemed to pass through similar stages of language development. Second, there was a growing realisation that systems of linguistic description like Bloomfield's did not provide an adequate basis for language learning because they focused 'too exclusively on knowledge of the formal linguistic system for its own sake, rather than on the way such knowledge is used to achieve communication' (McNamara, 2000, p. 14). The skills/components approach to testing seemed to be more concerned with language as a system than with learners' awareness of how to use that system in their lives, putting language to use as a means of communicating with other people.

As the behaviourist theory of language learning that supported it came under attack, the discrete-point testing of components was increasingly criticised. Applied linguists such as Corder (1971) and Selinker (1972) argued that it was unhelpful to view the language used by learners as being simply 'correct' or 'incorrect' as judged against the grammatical rules presented by language teachers. They argued that rather than forming language habits by repeating words and grammatical patterns, each learner developed a personal inter-language, shaping and refining their own (often subconscious) systems of grammatical rules as they developed their understanding of the new language. These rules did not necessarily conform either to the norms of their native language or to those of the target language.

A learner's inter-language could be more or less effective for communication even if it did not match the grammatical patterns assessed in discrete-point tests, which were based on a native speaker model. A discrete-point test of grammar might reveal whether or not a learner could apply a formal rule, but could not reveal the ways in which that learner might attempt to express an idea in everyday conversation or in a letter. As a diagnostic tool, such a test could only show whether or not the learner had taken up the taught patterns;
would reveal nothing about the rules that learners had formed for themselves, how far these might be from the intended target, or how effective they might be in real life communication.

**Example 7.4 Syntax: Recognition: Subject–verb agreement**

\[
\text{The [boys] strikes the car and runs.}
\]


In Example 7.4 from Lado (1961), an assessee may score a point by knowing that according to the subject–verb agreement rule the answer must be boy. However, the same assessee might misunderstand the meaning of the sentence — thinking it means that the boy starts the car and drives away — and might also be unable to tell the story of a boy hitting a car, based on a series of pictures. A second assessee might not be able to apply the subject–verb agreement rule, but might nevertheless understand the meaning and tell the story effectively. A discrete-point grammar test would reward the first assessee, even though he is less effective at using the language to communicate.

**Example 7.5 Phonetic contrast**

A. B.

Figure 7.3 Phonetic contrast — the test taker hears: 'Let me see the sheep'.


It was also pointed out by critics that it was very difficult to produce tests dealing with just one component of the language system, but not involving any other components. The classic *minimal pairs* test is intended to test the single component of phonetic discrimination. In Example 7.5 presented by Lado (1961), the assessee must distinguish between two sounds: the 'ee' sound /i:/ of sheep and the the 'i' sound /i/ of ship. But, as the critics pointed out, an assessee might fail to give a correct response for reasons unconnected with phonetic contrasts. Regardless of the sounds, an assessee might simply not know which word refers to the animal and which to the boat. In this case, the item would be testing vocabulary knowledge rather than phonetic discrimination.

There also seemed to be limits to the kinds of information that discrete-point tests could provide. In order to gain a full picture of whether or not a test taker was able to recognise the sound contrast in the ship–sheep example, presumably the test would need to include a wide variety of items testing recognition of the /i/ and /i:/ sounds in combination with different sounds and in different positions within the word — *is: ease; bin: bean; jelled: gelid* — and within different positions in sentences or utterances — *The sheep is over there, in the water. There is picture of a ship on the wall*. While this might be feasible in classroom-based assessment, where a teacher could reasonably spend a few minutes focusing on one sound contrast, it would certainly be impractical in a proficiency test intended to provide a general picture of the assessee's abilities within the very limited time usually available for testing.

Interpreting scores from discrete-point tests also seemed problematic. It was unclear how the many component parts should be put back together again to establish something meaningful about the test takers' ability to use the language as a whole. This problem involved both the content of the test (what balance of components would best represent overall language proficiency? should phonology and syntax be considered equally important?) and the meaning of the scores (what proportion of the questions on the test would the test taker need to be able answer correctly in order for users to infer that he or she knew enough language to cope with academic study or to practise as a doctor?).

When these objections were factored in, the scientific objectivity of the approach was fatally undermined. Certainly, the scoring processes were objective, but the process of compiling the test was not. Deciding on how many items should be included on the test to represent each of the components and deciding which component was actually represented by each item were matters for subjective judgement.
Integrative testing

**Task 7.5**

What is implied by the term ‘integrative testing’? How do you think this may have differed from discrete-point testing?

What kinds of task would you expect to find in an integrative language assessment?

In the other highly influential language assessment publication of 1961, Carroll agreed with Lado about the need to test ‘very specific items of language knowledge and skill judiciously sampled from the usually enormous pool of possible items’ (Carroll, 1961, p. 318). In other words, he also favoured a skills/components approach. But unlike Lado, he placed more emphasis on skills than on components. He was also less enthusiastic than Lado about the value of contrastive analysis. He felt that the issue for a proficiency test should be ‘how well the examinee is functioning in the target language, regardless of what his native language happens to be’ (Carroll, 1961, p. 319). He worried that the logic of contrastive analyses led towards the use of different tests of proficiency for learners from every different language background.

His solution was to recommend the use of what he called integrative tests. These were tests that could measure the rate and accuracy with which learners were able to bring together different aspects of language ability under realistic time constraints to produce or comprehend complete utterances or texts. Anticipating later developments in performance testing, he suggested that a starting point for tests for international students should be a detailed investigation of the kinds of language task that they might need to undertake at university and the level of performance needed to support academic success. Test developers would need to consider questions such as ‘what kinds of English mastery are required for the foreign student to comprehend reading matter in the several academic disciplines?’ or ‘what standard of proficiency in English pronunciation is required for foreign students to be understood by American students and teachers?’ (Carroll, 1961, p. 315).

A paper by Carroll et al. (1959) investigated a new set of techniques proposed by Taylor (1953) as a means of judging the suitability of written texts for children of different ages: cloze procedures. Cloze involves asking assesses to restore texts that have been partially erased. Points could be awarded either for providing precisely the word that had been erased from the source text (exact words scoring) or by providing any word that would fit in the context (semantically acceptable scoring). Carroll and his colleagues saw the potential of cloze procedures as integrative tests of language proficiency. A learner with low proficiency in a language might struggle to restore much of a text made up of relatively common words in simple sentences; a learner with high proficiency should be able to restore more missing material to texts composed of less common words in more complex sentences. Although the choice of scoring method apparently makes little difference when cloze procedures are used with native speakers, Carroll et al. (1959) saw the semantically acceptable scoring method as more appropriate for language learners.

**Task 7.6**

Look at Example 7.6. In what ways is this different from the examples of discrete-point tests in Example 7.2? What do you think might be the advantages and disadvantages of these cloze techniques?

---

**Example 7.6 Word and letter cloze techniques, based on Carroll et al. (1959)**

**Word cloze:** write a suitable word to fill each gap (1 to 6).

Penny Baxter lay awake beside the vast sleeping bulk (1) __ his wife. He was always wakeful on the full (2) __ . He had often wondered whether, with the light so (3) __ , men were not meant to go into their fields (4) __ labor. He would like to slip from his bed (5) __ perhaps cut down an oak for wood, or finish (6) __ hoeing that Jody had left undone.

Marjorie Kinnan Rawlings (1938) *The Yearling*

**Letter cloze:** choose a letter to fill each gap (7 to 12). An asterisk (*) represents a space between words.

(7) _SORBING*TH_ OPPED(8)_RAINI_ WAS*DAMP*A(9)_
(10)_BILTY* THE*TO(11)_ HIR(12)_*WE

*Answers*: Word cloze: (1) of (2) moon (3) bright (4) and (5) and (6) the
Letter cloze: (7) B (8) * (9) T (10) A (11) U (12) T

In Taylor’s (1953) work, cloze referred to the deletion of words at set intervals in a text – Carroll et al. (1959) deleted every tenth word in texts that were each 205 words long. However, it has also been used to refer to any kind of partial deletion – Carroll et al. (1959) also experimented with deleting the first, middle or last letter from sequences of 5, 7 or 11 letters (see Example 7.6) and with erasing sounds from recordings. A later variant that has been quite widely used in language tests is the C-test. This involves deleting the second half of every second word (Klein-Braley, 1985). However, teachers report finding this an annoying exercise.
Cloze procedures are based on the insight that language includes a good deal of redundancy: we can generally understand a message without attending to every word and often anticipate what we will hear or read next when listening or reading. Because elements tend to appear together with regularity, language users are able to draw on language habits to form an expectancy of what is likely to come next in a text. Proficient users of English will expect to see a vowel follow certain strings of consonants such as thr- or cl-. The nonsense words 'thronk' and 'clonk' are plausible in English, but the words 'thrko' and 'clnko' are not. In filling in a gap, learners may draw on awareness of syntax and lexis. They will expect a noun to follow an adjective in certain contexts and will be aware that some words are more likely to occur together than others – 'fast food' and 'a quick meal' are heard much more often in English than 'quick food' and 'a fast meal'.

Task 7.7
Pick a text from a textbook and make a cloze test by deleting every seventh word.

Make a C-test by deleting the second half of every second word, replacing each missing letter with an underscore.

What would these tests tell you about a learner's language abilities? Would any of the items be difficult to mark? Why?

Do the items only involve linguistic knowledge, or do they need knowledge of the topic to find the correct answer?

One of the attractions of integrative tasks such as the cloze procedure was that sets of passages ordered according to their difficulty for 'typical' language learners might form a rational scale for measuring competence in terms of the extent to which native performance in the language is achieved' (Carroll et al., 1959, p. 112). If a scale of this kind could be constructed, a new approach to reporting and interpreting scores would become possible. Carroll (1961, p. 321) suggested that scores might best be reported as probabilities – as in 'the chances are only three out of ten that an examinee with a given score level would be able to pass the freshman English course'.

This line of thinking was very influential in another major testing initiative of the period on which Carroll served as a consultant: the US Foreign Service Institute (FSI) Language Proficiency Ratings. This scheme was originally developed at the time of the Cold War to measure the language abilities of US government employees. In recognition of the need for government agents to be able to interact in foreign languages, a structured oral interview was used rather than discrete-point tests of language components.

Traditional speaking tests or 'oral examinations' had continued in use, largely unchanged in format, since the previous century. Tests typically consisted of a reading aloud phase, after which the examiner might ask a few questions based on the text that had been read. This would be followed by a brief conversation between the examiner and assesse. In the Cambridge examinations (CPE had been joined in 1939 by the Lower – subsequently First – Certificate in English), the examiner would award marks to the candidate for fluency, pronunciation and intonation. Certain numbers of marks might be associated with a certain standard as in 'Pass 9 / 20, Good 12 / 20, Very Good 14 / 20' (Roach, 1945, p. 7).

These speaking tests suffered from some obvious weaknesses:

- Because assessees needed to be tested individually by trained examiners, the tests were difficult and expensive to administer.
- Examiners could be inconsistent in their scoring, making the results unreliable.
- The scoring criteria were vague and did not seem to reflect linguistic theory.

The practical logistical limitations were inevitable, but perhaps overstated. Where the testing of speaking skills was considered important enough, the investment in examiners and support systems was made. Reliability, on the other hand, was more difficult to improve. The traditional approach to scoring both essays and orals relied on moderation. Experienced examiners were asked to check a sample of the marks given by junior examiners (of course, in the days before cheap, portable recording equipment, this was easier to do for essays than for oral examinations). They might also observe examiners while they conducted an oral, challenging and altering any marks that seemed too harsh or too lenient. Comparisons could be made between marks awarded for speaking and scores on other test components, or with results from the same school in previous years. Adjustments were made where individual examiners appeared overly harsh. Teachers familiar with the assessees might also be consulted. The system was hierarchical, with more experienced examiners assumed to be more accurate than novices and with the chief examiner as the ultimate authority.

Although these practices may have seemed satisfactory where small numbers of examiners were working in a familiar setting, there had long been concern that even experienced examiners varied in their interpretation of standards so that assessees could receive very different results depending on who was examining them. As secretary of the Cambridge examinations, Roach (1945) recognised that when tests were being given in over 50 different countries, it was difficult to be
confident that examiners working in all these different locations were all awarding marks in the same way and on the same basis. Roach experimented with the use of recordings of performances to train examiners in understanding the scoring system. He also introduced joint examining by two examiners – one sent out from Cambridge, the other a teacher working locally – to more effectively transmit and maintain the intended standards.

The third issue, the theoretical basis for the scoring criteria, was not addressed by Roach, perhaps because he was, in any case, sceptical of the benefits of criteria, arguing that ‘standards of impression marking cannot be well defined beforehand by written instructions’ (Roach, 1945, p. 38). This contrasts with the approach adopted for the FSI. Scoring was based on a very detailed rating scale used to focus the examiners’ attention during the interview on key features of test performance derived from (Bloomfieldian) linguistic theory: accent, comprehension, fluency, grammar and vocabulary (Wilds, 1975). The outcome of the interview was a rating on a scale that could be used to assess the employee’s ability in any language, ranging from 0 (no functional ability) up to 5 (equivalent to an educated native speaker).

In addition to rater training aimed at encouraging agreement on the interpretation of levels of performance, several other strategies were used to promote reliability. One was a degree of standardisation in the interview procedure itself, which followed a sequence of phases intended to ‘warm up’ the assessees with a few straightforward questions before increasing the level of challenge to find the point on the scale at which the assessees might not be able to function effectively. This was backed up by the collection of data on the extent to which raters agreed with each other in their ratings (see inter-rater reliability in Chapter 4).

The mastery learning movement

**Task 7.8**

What makes someone a successful language learner? Are some learners naturally better at learning languages than others?

What elements might you include on a test to find out whether someone would be able to quickly learn new languages?

Using graded language tasks such as reading passages or speaking activities of increasing difficulty to establish an individual’s overall proficiency in a language had obvious attractions for both teachers and testers. In language teaching it suggested a sound basis, mastery learning, for sequencing instruction. In proficiency testing it could be assumed that an assessee who was able to complete a cloze test or carry out a spoken activity at one level would also be able to successfully complete tasks at all of the levels below.

The idea of staged targets for language learning has a long history. Four hundred years ago, Comenius divided language learning into four ages, suggesting that only the child’s mother tongue and Latin as (the pan-European language of the educated) needed to be learned at the highest level:

- The first age is babbling infancy
- The second age is ripening boyhood
- The third age is matured youth
- The fourth age is vigorous manhood

Comenius’s language learning levels from the Great Didactic (Comenius, 1638)

The four stages corresponded to four language courses conceived by Comenius – the Vestibulum, Janua, Palatium and Thesaurus. The innovative features of the mastery learning movement were its technical rigour, its inclusivity and the formative role given to assessment. It was not only a matter of finding out whether learners could do the thing they had been taught, but of taking appropriate action to help those who were not yet able to do it.

Inclusivity was expressed through the concept of aptitude. For John Carroll (1963), everybody had the capacity to learn, but some were equipped to do this faster than others. Learning aptitude represented this variation between learners in the rate at which they could learn. Supporters of mastery learning suggested that the differences seen in achievement in the classroom could be reduced by increasing variation in the teaching (Bloom, 1968). In other words, if teaching could be better matched to learners’ individual aptitudes, it would bring improved results.

In traditional classrooms, all learners progressed through the material at the same speed, no matter whether they already knew what was being taught or were at a complete loss to understand any of it. Mastery learning brought greater variety: learners with high levels of aptitude would be allowed to progress quickly to more challenging tasks without being held up by slower learners. Those who struggled with basic concepts would not be left behind to sit in confusion as their classmates forged ahead, but could be given extra help until they could grasp the more basic material. To measure aptitude for language learning and to identify which learners were likely to make relatively rapid progress in language learning, Carroll, with Carl Sapon,
developed the Modern Language Aptitude Test (MLAT) (Carroll and Sapon, 1959), which measured features such as short-term memory and the ability to associate sounds and symbols.

In mastery learning approaches, the milestones used to chart progress along the continuum of proficiency are generally known as behavioural objectives. These objectives, often presented in the form of ‘Can Do’ statements, specify the observable behaviours that should result from successful learning. Ideally, these observable behaviours should be meaningful real world applications of the skill being taught (Block, 1971). Carroll (1971, p. 31) argued that ‘it is most essential ... to be able to state as exactly as possible what the learning task is, particularly its objectives, in testable form. That is, a teacher must be able to determine when a student has mastered a task to a satisfactory degree.’ Similar requirements continue to inform the design of objectives in language education (see Genesee and Upshur, 1996).

At the end of a unit of instruction, a test can be given, closely based on the specified objective. If teaching and learning have been successful, every learner should pass the test. Given such measurability, student progress may be tracked through periodic, formative testing. Feedback based on test performance can be used to guide extra instruction — using alternative methods — for learners who do not achieve the intended objectives (Bloom, 1968). This placed a new emphasis on testing (as well as teaching) by teachers, making assessment central to the classroom educational process.

Two approaches to score interpretation

Score interpretation

The history of assessment since ancient times has involved at least two contrasting approaches to the interpretation of scores. Bachman and Palmer (2010) call these relative and absolute score interpretations.

Relative score interpretations

One approach has been to treat performance as a competition: assesses are judged in relation to each other. In individual sporting contests, only the fastest athletes or strongest throwers are awarded medals. Similarly, in the competitive examinations used in ancient China and taken up in nineteenth-century Europe, only a few candidates would be chosen as worthy of a place working as government officials. Any number of people could enter for the examinations, but only a fixed number could be selected. If five jobs were on offer, the highest scoring five candidates would be chosen whether the number of entries was 5 or 500. In many traditional classrooms, each week the highest scoring student would be given a reward for coming ‘top of the class’; the lowest scoring student might be punished and humiliated.

In many individual competitive sports, such as archery, everyone has a set number of opportunities to shoot at the same target from precisely the same distance. Scoring is straightforward: different numbers of points are awarded for shots that hit different zones on the target. Similarly, in standardised examinations, everyone is given precisely the same tasks to perform and questions are designed with objective scoring in mind. However, it’s not always easy to tell from the results how well the competitors would perform under less controlled conditions. The learner in a class with the best control of grammar, the highest vocabulary test scores and great pronunciation is not always the one who is able to communicate most effectively with the foreign visitor.

Absolute score interpretations

An alternative approach to assessment has involved deciding whether or not the assessee is qualified to enjoy a certain status. Unlike the imperial examinations in China, in ancient Greece and Rome and in the mediaeval European universities qualifications were awarded when the assessee was pronounced by his teacher, or other competent judges, to be worthy to be called a scholar. In this kind of qualifying examination, there is no limit to the number of people who can pass. Modern examples of qualifying examinations that are often cited are the driving tests used in many countries. Any number of individuals may qualify as drivers, provided that they demonstrate their ability to control a car, show awareness of road conditions and display a good knowledge of the rules of the road.

Unlike competitive examinations, many qualifying examinations are performance assessments that involve the simulation of real life conditions. Because the focus is on carrying out realistic, often complex and somewhat unpredictable tasks, it is not easy to standardise conditions. Traffic conditions can be trickier in certain locations or at certain times of day, and different learners will have more knowledge of an essay topic or feel more comfortable talking to an older examiner than their classmates.

In complex performance assessments, instead of everyone being asked to perform the same task, individuals may have the freedom to choose different kinds of tasks with different levels of difficulty. Even when every assessee is given the same prompt, each has to interpret what is required and so plays a part in deciding what the task involves. Judgements about the assessee’s abilities will involve weighing the contribution of the qualities of each performance against the demands imposed by the task.
The psychometric approach to testing taken up by Lado (1961) was based on the measurement of differences in relation to an underlying trait (language ability) and in this respect was more compatible with the competitive model. The focus was on the individual's language ability and the intention was to find out what levels of language ability were to be expected in a particular population of people - what was 'normal' and the extent to which people differed from this norm. Tests were designed with the intention of spreading out learners as effectively as possible in order to differentiate between the more and less able. Through a process of standard setting, interpretations could then be established for different score levels. Expert judges would decide what score on the test would be sufficient to qualify an assee for different purposes.

Mastery learning seemed to require a different approach to measurement, closer to the qualifying model. The objective was not to identify a norm for a population and then to differentiate between those who possessed more or less of the attribute. Instead, learners who seemed to have reached a certain level of functional ability would be identified. They could then be administered a test to confirm their ability to carry out relevant tasks. For example, a teacher might decide that one of her students had enough ability to pass on simple telephone messages. The learner could be given a simulation task involving passing on messages in order to confirm the teacher's view. Different tests would have to be developed to represent different uses of language, presenting different levels of challenge. Scores should be interpreted in terms of achieving a criterion of communicative success in real world language use. This approach to interpreting scores is known as criterion-referenced measurement.

Pragmatic tests and the unitary competence hypothesis

In his book *Language Tests at School*, John Oller (1979) rejected discrete-point testing, arguing that valid language tests should be both integrative - testing language components in combination rather than attempting to separate them out - and pragmatic in the sense that they should engage the test taker's awareness of meanings and contextual constraints, as well as knowledge of linguistic forms. Along with the more integrative comprehension tests, oral interviews and essays disparaged by Lado, Oller was also enthusiastic about the potential of dictation and cloze techniques as integrative task types engaging multiple elements of language - form and meaning - at the same time. Where Carroll et al. (1959) had questioned the value of cloze tests on the grounds that they seemed to require too much knowledge of contextual information beyond the sentence, Oller saw this as one of the key attractions of the technique.

Oller (1979, p. 42) gives the following example of a cloze exercise:

In order to give correct responses, the learner must operate (1)____ on the basis of both immediate and long-range (2)____ constraints.

The correct answer to (1), on, can be deduced from the immediate context. In other words, the test taker could probably supply the correct answer even if only given: '____ the basis of'. On the other hand, (2) is more ambiguous. If the test taker is only provided with 'long-range ____ constraints', any number of words might fit the gap: 'Oller suggests missile, legal and leadership. However, the context - a text about language assessment and cloze procedures - rules out these possibilities and suggests words such as semantic, linguistic or, the original word, contextual.

At the time, some applied linguists such as Stephen Krashen suggested that the teaching of grammatical structures, the core of most language programmes, might be misguided. Second language acquisition research had revealed interesting consistencies in the order in which certain features of languages were acquired (regardless of the order in which they were presented). This suggested that there might be parallels between the way in which babies acquire the grammar of their first language without conscious effort and the way in which adults learn a second or additional language. The mental processing capacity of the brain might impose limits on what could be learned at different stages of development. Instead of endlessly repeating certain words, phrases and structures as in behaviourist classrooms, in Krashen and Terrell's (1983) 'natural approach', teachers should present learners with linguistic input at a level on the proficiency continuum just beyond what they were currently able to produce. The purpose of classroom assessment was therefore to discover what kinds of language learners could spontaneously produce and what kinds of text they could process.

Taking up Cronbach and Meehl's (1955) multi-trait, multi-method approach to validity (that tests of distinct abilities should yield different results while tests of the same ability should give similar results - see above), Oller noticed that the results of supposedly different test parts such as listening comprehension, oral interviews and grammar were often, in fact, very similar. In an interpretation that seemed to fit well with the evidence for common developmental sequences in language acquisition, he believed this was evidence that the different tests were all measuring essentially the same underlying ability: language proficiency. This suggestion, known as the unitary competence hypothesis, was briefly very influential, but lost credibility when Oller's interpretation of the statistical evidence was called into question.

Although the idea of a unitary competence was not ultimately upheld and the natural approach to language teaching failed to take
hold, the integrative approach to assessment did have a lasting effect. Cloze tests, in particular, proved very attractive to test providers because they were so practical to build. Unlike the skills/components approach, there was no need to invest in expensive test development and item writer training procedures because once a text was chosen, test construction was automatic. Unlike oral interviews or essay tests, there was no need to train raters because scoring was also relatively straightforward. However, contrary to Oller's hopes, further research suggested that cloze procedures did not often assess much more than the knowledge of grammatical structures and vocabulary found in their discrete-point predecessors.

The 'promised land' of communicative testing

The approaches to language assessment described up to this point share a view of language as an abstract system, deliberately distanced from the social contexts in which it was used. However, during the 1970s, following insights from the emerging field of socio-linguistics, there was a decisive shift in perspective. Hymes (1972) was particularly influential. Campbell and Wales (1970, p. 247) summed up the key issue: 'by far the most important linguistic ability has been omitted [from current linguistic theory] – the ability to produce or understand utterances which are not so much grammatical but, more important, appropriate to the context in which they are made'.

According to the situation, we make different choices about when to speak or write and for how long, about the words and grammatical forms we use and how we arrange them. For example, many people are likely to make rather different choices about how they describe their current boss according to whether they are being interviewed for a new job or relaxing at home with their friends and family. All aspects of language are affected, from the words and grammar we use to our rate of speech and our pitch, tone and accent. Socio-linguists began to explore variation in language and to look for the principles that might explain the different ways in which we communicate according to the aspects of the context we are in.

This shift in perspective filtered into the world of language assessment. In October 1980, a symposium was held in Lancaster in north-west England. A small group of language assessment specialists gathered to discuss a set of influential recent papers. The symposium led directly to the publication of a book summarising the discussions (Alderson and Hughes, 1981) and to the establishment of a newsletter that would later develop into the field's first dedicated journal, Language Testing. Dominating the agenda were papers rejecting Oller's evidence for a unitary language competence and promoting communicative or specific purposes approaches to assessment.

Wanting to introduce a more socio-linguistic perspective to language assessment, Cooper (1968) had made a distinction between the 'linguistic competence' dealt with in the skills/components model and what he called 'contextual competence'. Contextual competence involved knowing 'what to say, with whom, and when, where' (Cooper, 1968, p. 58) and this required its own distinct grammar, or set of rules. Although linguistic competence could be acquired independently of contextual competence, both were needed for effective communication. Cooper's distinction was taken up by Munby (1978, p. 17), who considered that 'Existing testing frameworks, concentrating as they do on linguistic competence, do not necessarily assess a person's communicative ability. Knowledge of the target language may not be sufficient for effective communication to take place in that language, and our ability to predict communicative competence depends upon the test content being based upon the specific communication requirements of the particular category of learner.'

The first paper considered at the symposium was Keith Morrow's (1979) manifesto for a communicative language testing. For Morrow, this was revolutionary and represented the ideal world or 'promised land' for testers. It would bring testing up to date and align it with the fashion for communicative teaching based on realistic interaction in the target language. The point was not to test knowledge of language systems, whether as discrete components or integrated whole, but the ability to use language functionally to carry out real world tasks: to understand and pass on ideas, to form and maintain relationships, to negotiate transactions.

Morrow was critical of Oller for failing to take account of the situations in which language is used. Believing that all uses of language reflected the same basic underlying ability, Oller (1979) had given only limited attention to issues of content, such as the kinds of text that should be employed for cloze procedures or dictations. He suggested, for example, that 'material intended for fifth grade geography students' (Oller, 1979, p. 365) might work well as a basis for cloze tests aimed at university-level students and that 'dictation at an appropriate rate, of kinds of material that learners are expected to cope with, is a promising way of investigating how well learners can handle a variety of school-related discourse processing tasks' (p. 269). He did not explore further how developers should identify these materials.

In contrast to Oller, advocates of communicative testing insisted on the authenticity of their test material, aiming to reproduce real life conditions in their tests (although, as Bachman, 1990, pointed out, they paid too much attention to the test content and not enough to how asseesees responded to this). Where Lado (1961) had used the
term 'integrated skills' to refer to listening or reading (as distinct from discrete elements such as vocabulary or intonation), Brendan Carroll (1980) used the same term to refer to the integration of modalities involved in carrying out real life tasks: listening and speaking integrated in conducting a conversation or discussion; reading, listening and writing along with graphics or other supporting visual material in preparing a simulated academic assignment.

The English Language Testing System (ELTS) (launched in 1979), on which Carroll worked, was designed, like the TOEFL test, to assess English language ability of international students, but took a radically different approach. In place of discrete tests of grammar and vocabulary, ELTS tested the use of reference material (in a component titled 'Study Skills'), included an integrated test of reading and writing, and offered test takers a choice of modules linked to different disciplinary areas.

The ELTS test (see Table 7.5) offered a choice between a Non-Academic test for vocational candidates, a General Academic option, or one of five modules covering broadly defined areas of study: Life Sciences; Social Studies; Physical Sciences; Technology; Medicine. In addition to three sections in their chosen subject area (M1), each candidate was required to take two common tests in the General (G) section.

Within the subject-specific part of the test, the three modules (M1, M2 and M3) were linked so that test takers wrote about a topic in the Writing paper (M2) that had been introduced in the Study Skills module (M1). The Interview (M3) would also involve a topic from M1.

### Task 7.9

Look at Example 7.7. How is this different from the other examples you have seen so far in this chapter? In what ways do you think these tasks more closely resemble real life language use?

### Table 7.5 The design of the ELTS test

<table>
<thead>
<tr>
<th>Subject specific modules</th>
<th>M1 Study Skills</th>
<th>M2 Writing</th>
<th>M3 Individual Interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>General modules</td>
<td>G1 General Reading</td>
<td>G2 General Listening</td>
<td></td>
</tr>
</tbody>
</table>

### Example 7.7 Sample items from ELTS

#### Listening (G2)

**SECTION 4**

**Listening to a Seminar**

Some evidence on spacing and spacing. What is the best order to do these? We'll try to show you how to conduct this experiment. Suppose you are members of the group and you are the chairman. You have to make sure that you decide on the order.

**TERMINAL: GROUP A**

**INTRODUCTORY COURSE**

<table>
<thead>
<tr>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.00</td>
<td>15.00</td>
<td>16.00</td>
<td>17.00</td>
<td>18.00</td>
</tr>
<tr>
<td>Lecture</td>
<td>Lecture</td>
<td>Lecture</td>
<td>Lecture</td>
<td>Lecture</td>
</tr>
<tr>
<td>Tutors</td>
<td>Tutors</td>
<td>Tutors</td>
<td>Tutors</td>
<td>Tutors</td>
</tr>
<tr>
<td>Special</td>
<td>Special</td>
<td>Special</td>
<td>Special</td>
<td>Special</td>
</tr>
</tbody>
</table>

**20.00**

**EXPOSURE**

After each group will be a chance for you to answer questions on your first draft. They are written in all my notes and you ask in groups, however, you ask them in pairs. Think carefully about these questions before answering.

#### Study Skills (M1)

**Table 7.4 Sample items from the ELTS Specimen Materials booklet**

**Source:** The British Council (1987) English Language Testing Service Specimen Materials booklet
Texts and task types in communicative tests were drawn as directly as possible from the target language use domain. When designing another new test of English for academic purposes, Weir (1983) took this principle even further than Brendan Carroll by using direct observation to answer the kinds of questions posed, but left unanswered, by John Carroll (1961) at the inception of the TOEFL. These included: 'What kinds of English mastery are required for the foreign student to comprehend reading matter in the several academic disciplines?' Using Munby's (1978) comprehensive framework for describing language use -- the communicative needs processor -- Weir carried out a form of job analysis, roaming university campuses, observing and characterising the English language requirements of the language use tasks that international students needed to carry out on their courses. The test that he produced at the end of this process, the Test of English for Educational Purposes (TEEP), required assesses to listen to a lecture and take notes, read an article on the same topic, then synthesise ideas from both in a short essay. In other words, they had to carry out, within the time limits imposed by the test, a typical academic study cycle.

Communicative and specific purpose-driven approaches to language testing were greeted rather sceptically by many researchers. Just as John Carroll (1961) had been concerned that contrastive linguistic analysis implied the need for a separate test for groups of learners from different language backgrounds, Alderson (1988) worried that as no two learners had identical needs, communicative testing implied a personal language test for every learner: not a practical proposition. Bachman (1990) argued that the communicative emphasis on successfully carrying out 'real world' tasks could not substitute for a clear definition of the knowledge, skills or abilities involved. The fact that an assessee succeeded in ordering lunch in a role play assessment would be rather meaningless unless it could be assumed that the abilities involved in doing so were really the same abilities required to order lunch in a restaurant or to carry out other similar activities. Finally, a language test, as Stevenson (1985) memorably expressed it, was 'a test not a tea party'. In other words, however carefully the assessment designer attempted to reconstruct real life conditions, the people involved understood that they were, in fact, participating in a test and so would not behave in the same way as in real life.

**Rapprochement: communicative language ability**

The Bachman model of communicative language ability (Bachman, 1990; Bachman and Palmer, 1996, 2010) built on earlier models suggested by Canale and Swain (1980) and Canale (1983), among others. The model brought together the task-focused communicative approach with the ability-focused skills/components tradition (see Chapters 5 and 6). In the model, communicative language ability involves language knowledge (see Figure 7.5) and other characteristics of the language learner, such as topic knowledge, strategic competence (planning, monitoring) and affect (motivation, anxiety). These interact with the characteristics of the task and setting (whether in a language assessment or real life context) to engender a performance. For Bachman and Palmer (2010), the language knowledge element is made up of organisational (grammatical and textual) and pragmatic (functional and socio-linguistic) components. Organisational competence enables the production of well-formed texts and sentences (or speech units), while pragmatic knowledge involves producing language that is appropriate to the context and to the user's goals for communication.

The pervasive influence of the Bachman model on subsequent language testing practice is evident in revisions made to both the more communicatively oriented ELTS test and the more structurally oriented TOEFL during the 1990s. In the revision of ELTS that eventually led to the 1995 version of its successor (IELTS -- the International English Language Testing System) some of its more experimental communicative aspects were jettisoned. The production of multiple versions to cater for learners from different disciplines was abandoned, along with the study skills modules and the thematic links between the reading and writing sections. The content was increasingly standardised through, for example, use of an interlocutor frame or partial script to control the part of the assessor in the Speaking test.

The revisions of the TOEFL test that culminated in the launch of the (internet-based) TOEFL iBT test in 2005 involved movement in the opposite direction. The structure and written expression and
vocabulary components were dropped, to be replaced by tests of writing and speaking that included greater integration with written and spoken input and output. The 2010 versions of the two tests (see Table 7.6) appear in many ways more like each other than their predecessors (Tables 7.4 and 7.5). Samples of current test materials can be found on their respective websites. Although there has been a degree of convergence between competing approaches, it would be misleading to suggest that there is now a consensus on how language abilities should be conceptualised for assessment purposes. While it is now generally agreed among language assessment specialists that language ability involves multiple components, the Bachman model has been criticised on the grounds that it fails to explain the relative contribution made by its various components or the dynamic ways in which they might interact with each other in communication (Widdowson, 2003).

The relationship between language ability and the contexts in which language is used has been a point of contention. For Bachman and Palmer (2010, p. 34), the components of the model interacted with, but were clearly distinguished from, other 'attributes of the individual language user', such as age, personality, etc., and characteristics of the language use situation, which included, among other things, the physical characteristics of the situation, the language use, and the interviewer's contribution to the interaction. In other words, different contexts imply different language abilities. The pragmatic competence that a learner has in university study is not necessarily the same as the pragmatic competence involved in working as a hotel receptionist. It cannot, therefore, be assumed that a learner who does well in a test of written language abilities will necessarily be as successful in a test of oral language abilities.

In the interactional view favoured by Chapelle (1998) and Chalhoub-Beck (2003), the contexts in which people use language, and the contexts in which they are tested, influence the nature of the test itself. For Bachman and Palmer (2010, p. 34), the components of the model interacted with, but were clearly distinguished from, other 'attributes of the individual language user'. The pragmatic competence that a learner has in university study is not necessarily the same as the pragmatic competence involved in working as a hotel receptionist. It cannot, therefore, be assumed that a learner who does well in a test of written language abilities will necessarily be as successful in a test of oral language abilities.

The question of whether a test is local to a specific context is a difficult one. Many learners, for example, are not confident in their ability to express their ideas in English in an academic setting, but are more comfortable in a social setting. The relationship between language ability and the contexts in which language is used has been a point of contention. For Bachman and Palmer (2010, p. 34), the components of the model interacted with, but were clearly distinguished from, other 'attributes of the individual language user'. The pragmatic competence that a learner has in university study is not necessarily the same as the pragmatic competence involved in working as a hotel receptionist. It cannot, therefore, be assumed that a learner who does well in a test of written language abilities will necessarily be as successful in a test of oral language abilities.

The question of whether a test is local to a specific context is a difficult one. Many learners, for example, are not confident in their ability to express their ideas in English in an academic setting, but are more comfortable in a social setting. The relationship between language ability and the contexts in which language is used has been a point of contention. For Bachman and Palmer (2010, p. 34), the components of the model interacted with, but were clearly distinguished from, other 'attributes of the individual language user'. The pragmatic competence that a learner has in university study is not necessarily the same as the pragmatic competence involved in working as a hotel receptionist. It cannot, therefore, be assumed that a learner who does well in a test of written language abilities will necessarily be as successful in a test of oral language abilities.

Table 7.6 The TOEFL IBT and 2012 IELTS tests compared

<table>
<thead>
<tr>
<th>Section</th>
<th>Time Limit</th>
<th>Questions</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>60-90</td>
<td>60 minutes</td>
<td>36-56 items</td>
</tr>
<tr>
<td></td>
<td>minutes</td>
<td></td>
<td>40 items</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Read 3 or 4 texts from academic texts and answer questions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Prose summary completion, table completion and multiple choice</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>34-51 items Chart completion and matching</td>
</tr>
<tr>
<td>Listening</td>
<td>60-90</td>
<td>30 minutes</td>
<td>40 items</td>
</tr>
<tr>
<td></td>
<td>minutes</td>
<td></td>
<td>Multiple choice, short answer, notes/summary/flow chart description,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Matching</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 conversations, 2 monologues</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>One of each is more social and one more academic in character -</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>discussion and lecture</td>
</tr>
<tr>
<td>Speaking</td>
<td>20 minutes</td>
<td>11-14 minutes</td>
<td>6 tasks</td>
</tr>
<tr>
<td></td>
<td>minutes</td>
<td></td>
<td>2 Independent tasks and 4 Integrated tasks</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Integrated tasks: test takers read a short passage, listen to related</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>materials, and then orally integrate the information in their own</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>words</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Independent tasks: test takers respond to familiar topic based on</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>own experience</td>
</tr>
<tr>
<td>Writing</td>
<td>60 minutes</td>
<td>60 minutes</td>
<td>2 tasks</td>
</tr>
<tr>
<td></td>
<td>minutes</td>
<td></td>
<td>2 tasks</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Write essay responses based on reading and listening tasks; support</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>an opinion in writing</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Task 1: test takers report on a diagram, table or data; Task 2: Test</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>test takers write an essay on a given topic based on their own</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>experience</td>
</tr>
</tbody>
</table>
Although most critics of the Bachman model have characterised it as overly cognitive, being concerned with individual mental processes rather than social and interactive aspects of language use, Weir et al. (2013) objected that the Bachman model, in fact, has little to say about how language is processed in the mind of the user. As a result, they suggested, the model offers no clear guidance on why one language use as overly cognitive, being concerned with individual mental processes to the development of Weir's (2005a) socio-cognitive framework, which accounts for task difficulty in terms of the demands imposed by different contextual characteristics on the assessees's cognitive resources (see Chapters 5 and 6).

Developments in formative assessment

By the 1960s, social and technological changes had given rise to new ways of thinking about education. It was realised that traditional patterns of 8 to 12 years of schooling, followed by a lifetime of employment in a familiar occupation, were unsustainable. Few children in industrialised countries could expect to spend their lives working in the fields or following other established occupations in the way that their parents or ancestors had. Instead, they could expect to experience rapid change even if they continued working for the same employer. Citizens would need to adapt to new industries and new forms of leisure. The concept of ‘permanent education’ (Council of Europe, 1970) or lifelong learning was suggested as a response to these new needs. People of all ages should be able to access education and training at many different points in their lives in order to make the most of new opportunities. More learning would take place outside formal school systems in the workplace or at home.

Since ancient times, education had been seen as a process of mastering a more or less fixed body of knowledge. Traditional systems of qualifications – school graduation certificates, university degrees, professional examinations – still embodied this classical model. They did not offer enough flexibility to reflect the explosion in new forms of knowledge that was going on. To provide for greater flexibility, while still communicating what learners had achieved, qualifications would need to become modular and learner-centred. In working towards and adding to their qualifications, individual learners would need to be able to choose different pathways, integrating knowledge, skills and abilities picked up in formal and nonformal settings. This implied that learners themselves would need to make informed choices about what and how they should study: learning how to learn as well as learning about academic subjects.

Lifelong learning and learner-centred education were compatible with mastery learning and graded objectives: working towards practical outcomes expressed in the form of Can Do statements. However, they also suggested that learners needed to become fully involved in setting objectives for their own learning and should take responsibility for keeping track of their own progress.

Following the lead of workplace education, by the 1990s, self-assessment, peer assessment and feedback were becoming established in language classrooms. Oskarsson (1980) and Brindley (1989) were pioneers of learner-centred language assessment, promoting the use of self-assessment checklists, portfolios and learning logs. In 1998, the publication of Black and William’s Inside the Black Box brought together evidence from over 650 studies showing that effective use of formative assessment could substantially improve the performance of school children across the curriculum. This gave a renewed impetus to learner-centred approaches in language assessment through the concept of assessment for learning.

Assessment for learning owes a good deal to the formative testing of the 1960s, but also differs from it in important ways. Poehner (2008) has pointed to a distinction that can be made between more interventionist and more interactionist approaches to assessment (not to be confused with interactional views of language ability). Mastery learning and formative testing had involved a straightforwardly interventionist approach. Learners might be given a pre-test – assessing how far they were already able to carry out a task. This would be carried out under controlled conditions. The teacher would look at the results and identify areas of strengths and weaknesses. He could then intervene to focus a unit of instruction on improving performance before administering a post-test to establish how much improvement had been made and whether learners were now able to carry out the task successfully.

Influenced by the work of Vygotsky (1986), and Bruner (1983), learner-centred assessment, or assessment for learning, introduced more exploratory interactionist approaches. The assessor could participate in the assessment, carrying out the tasks with the assessees collectively or individually. The rationale behind the approach is that if the assessees are able to improve on the initial performance when given hints or support, this indicates their immediate potential for learning. If the assessees fail to answer a question correctly, the assessor can step in with structured assistance or scaffolding. Through this process, the assessor learns and documents how much help is needed, and in what areas, to bring about improved performance. If the assessees are able to give a correct response, the assessor might ask follow-up questions to check whether they are able to apply their understanding to other similar problems, showing how well they have assimilated the concepts being assessed.

Differences between the two approaches are summarised in Table 7.7. For advocates of assessment for learning, the interactionist approach has two key advantages. First, it gives the students/assessees...
Table 7.7 Interventionist and interactionist assessment for learning

<table>
<thead>
<tr>
<th></th>
<th>Interventionist</th>
<th>Interactionist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questions addressed</td>
<td>How much has the assessee already learned?</td>
<td>How does this person learn?</td>
</tr>
<tr>
<td></td>
<td>What can he/she do? What can’t he/she do?</td>
<td>Under what conditions and by how much can performance be improved?</td>
</tr>
<tr>
<td></td>
<td>How does this person’s performance compare with other people’s?</td>
<td>What are the main obstacles to a better level of performance?</td>
</tr>
<tr>
<td>Assessment focus</td>
<td>Retrospective: products of past experiences</td>
<td>Prospective: processes involved in learning and prospects for improvement</td>
</tr>
<tr>
<td></td>
<td>Standardised: same for everybody</td>
<td>Individualised: responsive to what is discovered during the assessment</td>
</tr>
<tr>
<td>Assessment conditions</td>
<td>Independent</td>
<td>Supported</td>
</tr>
<tr>
<td></td>
<td>One opportunity to perform</td>
<td>Repeated opportunities to perform</td>
</tr>
<tr>
<td>Assesses performance</td>
<td>Presents tasks</td>
<td>Presents tasks</td>
</tr>
<tr>
<td></td>
<td>Records responses</td>
<td>Identifies obstacles</td>
</tr>
<tr>
<td></td>
<td>Disengaged</td>
<td>Engaged: mediates tasks, provides feedback</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Promotes learning</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Scores as estimate of current state of knowledge, level of skill or ability</td>
<td>Descriptions of learning potential: identification of obstacles to learning and of what is possible with intervention of the assessor or other expert</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identification of obstacles to learning and performance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Suggestions on ways of helping learners to overcome these obstacles</td>
</tr>
<tr>
<td>Interpretation of results</td>
<td>Identification of outcomes of learning, strengths/weaknesses to inform future teaching</td>
<td>Identification of outcomes of learning, strengths/weaknesses to inform future teaching</td>
</tr>
<tr>
<td></td>
<td>Identification of need for intervention</td>
<td>Identification of need for intervention</td>
</tr>
</tbody>
</table>


The process of mediation constructs for the learners or assesses a zone of proximal development: the distance between what they are able to do with support from others and what they are able to do independently. This indicates where there is immediate scope for them to learn. Poehner (2008) gave an example of two learners who made an error with verb tenses while speaking. One was stopped by the teacher and was immediately able to correct herself without further help; the other was only able to give the correct form when it was modelled for her. Although both assesses made errors, the learner who needed more scaffolding was clearly much further from being able to use the appropriate forms independently. The rich understanding of a learner’s state of development that can be obtained is, it is argued, potentially far more informative for users than traditional proficiency and achievement test scores.

Assessment for learning advocates are critical of teaching that prioritises interventionist over interactionist assessment. Black and William (1998), for example, suggested that teachers in many countries lack effective training in assessment and so base their practice on the examples provided by external tests. They identified a number of weaknesses commonly found in teachers’ use of assessment. Many of these weaknesses involve assessment practices that would be equally unacceptable in external testing programmes:

- Lack of collaboration and review: methods used by teachers are not shared and reviewed by colleagues, so good practices fail to spread and poor practices are not noticed and corrected.
- Construct under-representation: teacher-made assessments often focus narrowly on things that are easy to assess and neglect other areas (e.g., teachers find it easier to give grammar and vocabulary tests and may avoid assessing speaking skills because this is more challenging both to manage and to score).
- Testing as ceremonial: assessment is often treated as a bureaucratic chore or ritual with the objective of satisfying the expectations of managers or parents rather than as a way of collecting meaningful information about learners.
- Failure to use assessment information: teachers often ignore assessment records passed on by their learners’ previous teachers.

Others involve practices that could be considered acceptable in some external testing, but that are not so appropriate for the language classroom:

- Normative score interpretations: an overemphasis on competition and grading that can eventually persuade lower scoring learners that they are unable to learn.
- Lack of feedback: marks and comments made on learners’ work that do not make it clear what the learners need to do to improve.

a real and immediate opportunity to learn. Because they receive feedback and have opportunities to correct themselves, they can learn from the experience of assessment in a way that is not possible in assessments given under traditional test conditions. Second, it is claimed, an interactionist approach to assessment not only provides better tools for teaching and learning, but is also more revealing than traditional tests because it can inform the user about how far each learner will have to develop in order to perform independently.
• Speededness: teachers are often pressured to complete a given number of units of work within a set period rather than moving at a rate that is more appropriate to their learners’ needs.

Among those who embrace the concept of assessment for learning, a distinction can be made between people who take a more cognitive view of language learning, which is compatible with the Bachman model and socio-cognitive approach, and those who favour a more socio-cultural analysis. Like the early advocates of mastery learning, the cognitivists accept the dominant model of language learning which involves stages of acquisition and growth in individual functional abilities. They see assessment for learning as a means of helping learners to progress along an imagined scale of ability from beginner to mastery. In this view, standardised testing and classroom-based assessment have the potential to complement each other within a coherent system.

Some advocates of socio-cultural theory, in contrast, reject both the measurement metaphor of scales of ability and the view of languages as abstract systems of rules. In this vein, Poehner and Lantolf (2005) and Poehner (2008) argue in favour of a form of dynamic assessment which is opposed to traditional testing (or static assessment as they would characterise it). In this perspective, language learning does not happen in the mind of the learner, progressing along a predetermined pathway of individual development, but is a social process of assimilation into a specific local set of cultural practices and relationships. In contrast to the Bachman model and going well beyond the interactionist standpoint, language is seen to be inseparable from context. Development occurs as new forms of interaction are gradually taken up and internalised by the learner. The learner becomes an increasingly active participant in a particular community, with the teacher providing mediation between the learner and the target culture. Mediation, which characterises dynamic assessment, is contrasted with feedback. Poehner (2008) argues that feedback is focused only on successful task completion, while mediation is focused on the development of underlying abilities. However, as Leung (2007) notes, this is not a characterisation that would be acknowledged by many advocates of assessment for learning.

From the socio-cultural point of view, current frameworks and scales are simply an expression of one (dominant) set of values (Leung and Lewkowicz, 2006): values that can be contested and that are subject to change. One example of shifting standards has been the place of the ‘educated native speaker’ in rating scales. In the FSI scales, the highest level learners were defined as ‘functionally equivalent to that of a highly articulate well educated native speaker’ (Lowe, 1982). Over time, the ideal of native-speaker competence as the ultimate objective of language learning has been rejected. It is recognised that non-native language users can communicate effectively without being mistaken for native speakers and may prefer to retain a distinctive identity. References to native-like performance in the highest points on proficiency scales have now generally been replaced by formulations such as ‘precision, appropriateness and ease’ in language use (Council of Europe, 2001). This suggests for socio-cultural theorists that assessment constructs such as levels of language learning and proficiency are not predetermined properties of the human mind, fixed in the same way as physical properties, but that they shift according to what is most valued at the time.

Assessment wars: teacher assessment versus external tests

Standardised testing and classroom-based assessment (especially assessment of a more interactionist nature) are clearly very different. However, these differences are not always seen to be problematic, provided that each approach is reserved for a different purpose. Many proponents of interactionist approaches to assessment are prepared to accept a need for standardised proficiency testing and are pleased to show that increased use of formative assessment in the classroom can lead to improved performance on public tests (Black and William, 1998). Equally, many advocates of standardised testing acknowledge the benefits of interactionist assessment by teachers for effective learning, but might not believe that it has much value for decision-making beyond the individual classroom. Some writers on language testing go so far as to exclude less formal classroom procedures from their definitions of assessment (see, for example, Bachman and Palmer, 2010).

It is in the territory of summative achievement assessment or assessment of learning that there is greatest scope for conflict. This is an area where both teacher assessment and standardised testing may seem appropriate. More test-oriented and more learner-centred approaches are each seen to encroach on the domain of the other.

From the learner-centred point of view, when standardised tests are used – or ‘dehumanizing and oppressive pseudo-assessments’ as van Lier (2004, p. 29) refers to them, hinting at the emotional tone of the debate – they can come to dominate the curriculum, taking control over the content and pacing of instruction away from the teacher and encouraging teaching to become more test-like (see the section on ‘Washback’ in Chapter 4). It may appear to teachers that that the knowledge of learners that they have built up over an extended period is discarded in favour of the limited snapshots of learners’ abilities provided by tests. More formative applications of assessment are crowded out by the need to cover the content that will be tested.

On the other hand, from the psychometric point of view, classroom-based assessment involves a variability of conditions and procedures
that is incompatible with the goal of objective measurement. Relying on teacher judgements for high stakes decisions seems like a regression to a time before testing: a time when unsystematic procedures and biased judgements went unchallenged.

Some suggest that teacher assessment or classroom-based assessment, at least for summative purposes, should be judged against the same quality standards as external tests. Brown and Hudson (1998), for example, argued that the validity of inferences made on the basis of teacher assessments should be demonstrated in similar ways, using similar techniques, as for standardised tests. Others have argued that the tools used for judging the quality of tests are not readily transferable to more complex performance assessments (Moss, 1994). Interactionist assessments certainly cannot be reliable according to the established indices because conditions are flexible and everyone may finally give a successful response to each item. As a result, numeric scores based on the proportion of correct responses cannot be used to show which learners have been more and which less successful, even if teachers feel that they have gained powerful insights into their students’ learning (Poehner, 2008).

The advantages of summative assessment by teachers include the greater variety of tasks that it is feasible for teachers to employ over a period of time; greater potential for integration between summative and formative assessment within the curriculum; insights into learning processes, as well as outcomes; potential for professional development with better understanding of assessment issues; and greater flexibility to experiment with innovative task types. One major concern is that resources may not be made available to support the significant work involved in training, building shared standards, generating materials and moderating. Because of this, results may be inconsistent. However, if sufficient investment is made, teacher assessment of learning can be reliable. Although it may not be possible or desirable to standardise teacher assessments, Harlen (2007) collected research evidence suggesting that if scoring criteria are clearly specified, training provided and teachers’ work moderated, the results over an extended period are at least as reliable as those from standardised tests.

Another concern is that when teachers are asked to take on responsibility for summative assessment, this can place them in the contradictory and potentially corrupting position of being held accountable for training learners to succeed and at the same time judging whether or not they have achieved success (Brindley, 1998; Teasdale and Leung, 2000). Many of the same problems – such as narrowing of the curriculum – have been reported with high-stakes teacher assessment systems as with external tests. Such problems seem to be most acute when objectives and standards are imposed at a policy level without sufficient involvement of teachers and others

affected by the decision-making process (Rea-Dickins, 2001). This has led to prescriptions for more ‘democratic’ systems of assessment that involve a wider range of people at the design stage and that rely on multiple sources of evidence, including both classroom-based assessment and external tests (Shohamy, 2001).

Content standards and frameworks are often used as a means of lending coherence to national and international systems of language education and can provide a link between curricula, classroom-based assessment and examinations. Examples include the Canadian Language Benchmarks (Pawlikowska-Smith, 2000) and the US TESOL PreK-12 English Language Proficiency Standards (TESOL, 2006); but perhaps the best known internationally is the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). This is intended to aid communication between language educators working in different sectors and in different countries. Language assessments can be mapped to the overall scheme in terms of the content (what uses of language and what contexts of use does the assessment relate to?) and in terms of level. The framework partitions classroom language learning into three broad levels – A: Basic User; B: Independent User; and C: Proficient User – each of which can be subdivided to make finer distinctions. To aid the process of developing tests, the Council of Europe has published a Manual for Test Development and Examining for use with the CEFR (Council of Europe, 2011) and to help the developers to relate their assessments to the framework there is a Manual for Relating Language Examinations to the CEFR (Council of Europe, 2009).

The CEFR has proved popular with policy-makers and others (such as publishers and test providers) seeking to communicate about standards of language ability. On the other hand, the framework has been criticised as being too vague to use as a basis for test specifications (Weir, 2005b) and its role in policy-making has been called into question (Fulcher, 2004). One concern is that it has sometimes proved too easy to ignore the horizontal or content dimension of the framework when considering levels. If, for example, it is decided on the basis of a job analysis that the level of ability in a language required for someone to work as an air traffic controller equates to B2 on the CEFR, it does not follow that anyone assessed with a B2 level will have a suitable profile of language abilities. Standard validity arguments must still apply. Regardless of the CEFR, the assessment content must be shown to justify the inference that the assessee has the language abilities needed for the work. The content must be relevant to the target language use domain and the assessment results must be of adequate quality to justify decision-making.

However controversial its use may sometimes be, the CEFR has certainly raised the profile of assessment issues across Europe and around
the world. Its greatest contribution may be the process of reflection that it promotes. Each section of the framework concludes with an invitation: ‘Users of the Framework may wish to consider and where appropriate promote. Each section of the framework concludes with an invitation: ‘Users of the Framework may wish to consider and where appropriate promote. Each section of the framework concludes with an invitation: ‘Users of the Framework may wish to consider and where appropriate tests, DIF detection requires data from large numbers of assesses and therefore is limited to relatively large testing programmes. If DIF can be traced to construct irrelevant factors, it is evidence of unwanted bias. For example, McNamara and Roever (2006) recounted a study by Takala and Kaftandjieva (2000) that compared performance of male and female assesses on a 40-item vocabulary component of the Finnish Foreign Language Certificate Examination. The researchers found that males tended to perform better on items such as grease, rust and rookie, but females performed better on ache, turn grey and jelly (McNamara and Roever, 2006, p. 111). Unfortunately, DIF is often difficult to interpret. The words listed above might seem to reflect the roles of men and women in Finnish society, but it was less clear why women were more likely to know the word plot and men to know the word association. It is often unclear how far detecting instances of DIF implies bias or unfairness. If males are more likely to know the word rust, should it be removed from the test, or is its DIF simply a fair reflection of the realities of language use?

Related to the concern for fairness, there has been a growing interest in the profession of language assessment itself, its history, the identity of its members and their roles and social responsibilities. This has been reflected in the publication of histories such as Spolsky (1995), Davies (2008a, 2008b) and Weir (2013) and by guidelines on ethical questions: the International Language Testing Association’s (ILTA) Code of Ethics (2000) and Guidelines for Practice (2007) are both available for download from the ILTA website at www.iltanline.com.

The ILTA Code of Ethics sets out nine principles covering the responsibilities of assessment professionals towards assessees, the profession and society more generally. For example, Principle 7 states that ‘Language testers in their societal roles shall strive to improve the quality of language testing, assessment and teaching services, promote the just allocation of those services and contribute to the education of society regarding language learning and language proficiency’ (International Language Testing Association, 2000). The Guidelines for Practice outline the obligations of test designers, producers, administrators and users and the rights and responsibilities of test takers. They include, for example, requirements for designers to clearly state the intended purpose of their tests, for the accuracy of scoring to be checked and reported, and for users to be aware of the limitations of assessments and to ensure the relevance of the construct to the decisions they intend to make.

The ILTA Code of Ethics and Guidelines are not policed – there is no committee to check that they are being followed by members of the association – and there are no formal sanctions when they are breached. Although McNamara and Roever (2006) and others have questioned their value, criticising their brevity and vagueness, they do provide language assessment professionals with guidance on ethical behaviour and can support arguments for good practice and quality management within organisations. They serve to promote the sense of language assessment as a shared, global endeavour with its own distinct identity and underline the importance of making assessments as good as they can be.

If the ILTA Code of Ethics and Guidelines represent a growing awareness of the social responsibilities of language testers, some
commentators have gone much further in questioning the role of tests and assessments as instruments of coercion and social control. Shohamy (2001), reflecting the development of critical applied linguistics (Pennycook, 2001) and drawing on the work of social theorists such as Foucault (1975), called for a critical language testing that would identify and challenge the political values and assumptions behind assessment practices. The apparent technical objectivity of tests can conceal a political agenda, and we all need to ask searching questions about whose interests are really served when assessments are used.

As an example of the abuse of the power of language assessment, McNamara and Roever (2006) pointed to tests that are ostensibly intended to improve the employment prospects of immigrants, but that have been used to adjust the numbers of immigrants to a country or to block immigration by people from certain backgrounds. Kim and Elder (2009) questioned the fairness of the International Civil Aviation Organization (ICAO) policy on testing the English language proficiency of pilots and air traffic controllers. The policy places the onus on non-native users to bring their English to an appropriate standard, but fails to address the part sometimes played by the poor communication skills of native speakers when potentially dangerous incidents occur. Similar criticisms have been made of the use of tests of academic language ability. Learners are obliged to improve their language skills to gain entry, but institutions are not required to provide more accessible courses and materials adapted to the needs of language learners. Alderson (2009) brought together a range of examples of the ways in which institutional and national politics can affect how language assessments are used.

**Higher, wider, deeper: some emerging trends in language assessment**

Technology has always played an important part in language assessment, most visibly in delivery and scoring. It is now far easier than ever before to convey assessment material to the assessees and capture performances for scoring. Current developments, such as the transition from desktop computers to handheld devices, are certain to impact upon the nature of language assessment in the future. Most large-scale international tests and many classroom-based assessments are now delivered by computer, with test material appearing on screen and writing with pen and paper replaced by word processing.

The effects on performance and the equivalence of results when tests are delivered by computer in the form of recorded questions in a speaking test or on screen in a reading test have now been researched in some detail (see Chapelle and Douglas, 2006). On the other hand, the potential of computer-based test delivery to support innovation in the content of assessments is only beginning to be explored. Computer-based assessments are generally conservative and tend to reflect what was already possible in paper-based (or face-to-face) assessments. Computer-based tests of literacy skills, for example, still tend to give priority to text, but we can expect to see assessments that involve processing more multi-media and game-like material.

Scoring tests by machine has a long history, particularly in connection with selected response items. Advances in automated speech recognition and computational linguistics have now made it possible for machines to score written and even spoken performance, especially when responses are more guided or controlled (Xi, 2012) (see Chapter 4). The prospect is for more flexible automated scoring systems to emerge: systems that are capable of rating language elicited through different prompts without having to be trained for each individual task.

Computer scoring has the advantage of providing instant results and this feature can be exploited to create individualised assessments. **Computer adaptive tests (CATs)** are sensitive to the ability level of the assesse. If the assesse gives an incorrect response to the first item on a CAT, the computer will select an easier item next. If the first response is correct, the second question will be more difficult. Proceeding in this way, the computer selects material that increasingly closely matches the assesse's level of ability. CATs can be much more efficient than traditional linear tests because they can arrive at an accurate estimate of ability with relatively few items. Low-level assessees are not faced with 95 questions they cannot answer correctly in addition to the 5 that they can. On the other hand, the computer needs to have enough high-quality pretested material of known difficulty to cope with different patterns of responses and this makes CATs relatively expensive to produce. CATs have been available for over a quarter of a century, but have a chequered history in standardised testing. Perhaps because they give test takers little sense of how well they are performing, they have not always proved as popular as test providers hoped they would be. A persuasive argument for the fairness of standardised tests is that the challenge is almost identical for everyone. Individualising content risks throwing away this source of public sympathy for testing. However, new thinking about assessment suggests a new role for CATs. Staged support or scaffolding and feedback can also be programmed into computer-based assessments. An incorrect response to a short answer question may be used to trigger the presentation of multiple choice options or other hints. Errors in extended responses can be highlighted and advice provided for the writer.

By combining feedback mechanisms of this kind with computer adaptivity and detailed record keeping, **intelligent tutoring systems** (Wenger, 1987) involving complex virtual reality simulations can be developed. These will track learner progress and select appropriate
interpretations, Brown and Hudson (2002) and Brown (2005) provided
long been discussed, this application of technology is still in its infancy
instructional material as the learner moves forwards. Although it has
it has the potential to radically transform language education, seamlessly
An established view of language assessment sees it as combining
applied linguistics and measurement theory (Bachman, 2004). Reflecting
the latter, language assessment validation research has involved the use
of increasingly sophisticated computer-based analysis tools. Bachman outlines the use of statistics in describing test scores, improving test material and evaluating test use. Tools such as classical test theory, Rasch measurement, item response theory and generalisability theory are routinely used by test providers in evaluating test material and identifying the extent and possible sources of measurement error.

While classical test theory is founded on norm-referenced score interpretations, Brown and Hudson (2002) and Brown (2005) provided useful accounts of alternative statistics available for criterion-referenced testing. Other statistical tools such as factor analysis, regression analysis and structural equation modelling are regularly reported in the research literature because they provide evidence of relationships and patterns in test responses that suggest how tests or test components reflect underlying abilities. More complex performance assessments demand more complex measurement and the value of innovative techniques for analysing language assessment results is a recurrent theme in validation research.

As well as being used to investigate the outcomes of assessments, technology is increasingly being used to explore the processes involved when learners engage with assessment material. Video recordings, eye-trackers (which record where assesses focus their attention) and keystroke loggers (which record processes of writing and revision) have been used to explore how assesses tackle assessment tasks (Strömqvist et al., 2006; Winke, 2013). These provide insights into cognitive processing and supplement the information available from verbal reports by assesses on how they have responded (Green, 1998).

Corpora are electronic databases that sample written or spoken language from specified contexts in a form that makes it amenable to detailed linguistic analysis. Corpora that sample defined contexts of language use (medical language, legal language, etc.) can be used in the design of tests of languages for specific purposes (Barker, 2013). On the other hand, samples taken from test performances can be used in the development of learner corpora to tell us more about the nature of the language that is actually produced in response to assessment tasks. Study of learner corpora reveals the linguistic forms that are used and the errors made by learners from different backgrounds at different levels of ability. As part of a wider programme of research called English Profile, Hawkins and Filipovic (2012) explored the language used by assesses at different levels of the Common European Framework of Reference for Languages (CEFR) in completing assessment tasks. They derived from this what they term critical features of learner language that distinguished most efficiently between higher- and lower-level learners.

Weir et al. (2013) reported that at its launch just three assesses took the Cambridge Certificate of Proficiency in English (and all three failed). Today hundreds of thousands or even millions of people take the larger national and international tests each year and the expansion shows no signs of abating. One aspect of the recent growth in numbers has been diversification in the assessments on offer. The growing numbers of assesses make it more practical to provide assessments that have a more specific focus. Tests of languages for general business or academic purposes have been joined by tests for engineers, lawyers, accountants and other professions that involve international cooperation. Reflecting changing educational policies around the globe, large numbers of younger children are now taking language assessments. This has led to the launch of new international assessments for young learners and the appearance of a specialist literature relating to this area (see, for example, McKay, 2005).

The involvement of teachers in high-stakes summative assessment has also expanded. Even traditionally examination-driven school systems such as Singapore and Hong Kong have begun to incorporate school-based assessments into their educational systems (Davison and Hamp-Lyons, 2010). With growing awareness of the social responsibilities of the language assessment profession, this has led to calls for professionals to work together to improve the assessment literacy of teachers, policy-makers and others and to provide better public information on assessment issues. The closed world of assessment expertise is being opened up to public scrutiny. In some countries, this has led increasingly to legal challenges to assessment practices and score outcomes – another powerful incentive for assessment providers to take their social responsibilities seriously (Fulcher and Bamford, 1996).

As reflected in the account presented here, language assessment as an academic field and as a professional endeavour has its institutional origins in the US, Britain and a few other countries, predominantly English speaking (Spolsky, 1995). It has tended to centre on the assessment of English as a foreign language internationally, English as a second or additional language for migrant groups, and on modern foreign languages taught in English-speaking countries. Today, its influence on institutional practice may remain limited, but it has spread to all parts of the globe and centres of language assessment expertise can be found in every region and in the assessment of a wide range of languages. Although the established centres continue to thrive, it seems likely that as the profession continues to expand, innovative techniques in assessment design and quality assurance will emerge from new sources.
Last words

We have discovered a great deal over the past century about how languages are learned. New fields of research such as neuroscience and computational linguistics are opening up new insights all the time. For all that, our understanding of the intricacies of the process is still very limited. The quality of the information we can gain from language assessments is therefore unavoidably partial and uncertain. This makes it all the more important that we take full advantage of the resources we do have at our disposal – the PRICE elements of planning, reflecting, improving and working cooperatively – to generate the best evidence we can to inform our decisions.

There is a pressing need for a shift in attitudes towards assessment in language education: a shift that is only beginning to happen. Assessment is, as I hope this book makes clear, an indispensable ingredient in effective teaching and learning. Unfortunately, in teacher training it is often treated as a marginal component, poorly integrated with other modules and often relegated, like an afterthought, to the end of the course. As a result, teachers often know and care far less about assessment than about other aspects of teaching. Many feel poorly equipped to carry it out. Language textbooks do not generally provide them with the help they need, failing to match assessment activities with learning objectives in useful ways. Books that are designed to prepare learners for external tests have long been popular, but they typically offer little more than practice in test taking, with inadequate guidance on how to cultivate and develop learners’ skills and understanding in ways that match the objectives of the test.

Teachers have much to learn from the quality standards employed in external tests and can benefit from learning about effective assessment design. Tests themselves can often provide information that is useful for teachers and learners: information that is too often under-exploited. On the other hand, external tests with their rigid, standardised conditions and delayed feedback do not provide a good model for classroom activity. Alternative forms of assessment are often more engaging and may be more useful and more enlightening about learning processes. Exploring the alternatives open to them is important to teachers’ professional development.

Ultimately, the value of assessments depends on the people involved in designing, building, administering and scoring them and those who interpret the results. Good practice in language assessment, finding the most appropriate tools for each purpose, requires commitment and expertise. I hope that this book will encourage readers who are involved in language education to make that commitment and seek to learn more.

8 Commentary on selected tasks

Chapter 2

Task 2.1

1 How long is it likely to take this person to learn to speak at an advanced level? This information would be useful to the person him/herself, as well as to teachers and to anyone who is thinking of paying for the language classes or study abroad that might be needed. Obtaining that information could involve an assessment of language learning aptitude. It might be equally helpful to look at charts of how long it typically takes learners to reach an advanced level when studying this language in similar circumstances.

2 What sounds does this learner find it most difficult to produce? This will probably be of most interest to the learner and his/her teachers. This is a diagnostic purpose.

3 Is this learner going to benefit more from an elementary- or an intermediate-level language class? This would interest school administrators who need to put learners into suitable classes: placement purposes.

4 Has this student learned enough to progress to studying at the next level? Again, this is likely to be of most interest to school management, teachers and learners. The purpose of any assessment would be to focus on achievement (and the emphasis is clearly summative rather than formative, although results could also be used to focus instruction on areas of weakness – formative diagnosis).

5 Has this student learned the definitions of the 20 words assigned as homework? This is of interest to the teacher and student (and perhaps parents or others with an interest in the student’s learning). This involves assessment of achievement (which is potentially formative if used to guide further study).