CHAPTER ONE

Introduction

Speaking skills are an important part of the curriculum in language teaching, and this makes them an important object of assessment as well. Assessing speaking is challenging, however, because there are so many factors that influence our impression of how well someone can speak a language, and because we expect test scores to be accurate, just and appropriate for our purpose. This is a tall order, and in different contexts teachers and testers have tried to achieve all this through a range of different procedures. Let us consider some scenarios of testing speaking.

> Scenario 1
> *There are two examinees and two testers in the testing room. Both examinees have four pictures in front of them, and they are constructing a story together. At the end of their story, one of the testers asks them a few questions and then closes the discussion off, says goodbye to the examinees, and stops the tape recorder. After the examinees leave, the testers quickly mark their assessments on a form and then have a brief discussion about the strongest and weakest features of each performance. One examinee had a strong accent but was talkative and used quite a broad range of vocabulary; the other was not as talkative, but very accurate. They are both given the same score.*

This is the oral part of a communicative language assessment battery, mostly taken by young people who have been learning a foreign language at school and possibly taking extra classes as one of their hobbies. The certificates are meant to provide fairly generic proof of level of achievement. They are not required by any school as such, but those who have them are exempted from initial language courses at several universities

and vocational colleges. This may partly explain why the test is popular among young people.

Scenario 2
The language laboratory is filled with the sound of twelve people talking at the same time. A few of them stop speaking, and soon the rest follow suit. During the silence, all the examinees are looking at their booklets and listening to a voice on their headphones. Some make notes in their booklets; others stare straight ahead and concentrate. Then they start again. Their voices go up and down; some make gestures with their hands. The examinees' turns come to an end again and another task-giving section begins in their headphones. The test supervisor follows the progress of the session at the front of the room. At the end of the session, the examinees leave the lab and the supervisor collects back the test booklets and starts the transfer of performances from the booths to the central system.

To enable the test session in Scenario 2 to run as expected, many steps of planning, preparation and training were required. The starting point was a definition of test purpose, after which the developers specified what they want to test and created tasks and assessment criteria to test it. A range of tasks were scripted and recorded, and a test tape was compiled with instructions, tasks, and pauses for answering. The test was then trialled to check that the tasks work and the response times are appropriate. The rating procedures were also tested. Since the system for administering the test was already set up, the test was then introduced to the public. The scores are used, among other things, to grant licenses to immigrating professionals to practise their profession in their new country.

Scenario 3
Four students are sitting in a supposed office of a paper mill. Two of them are acting as hosts, and the two others are guests. One of the hosts is explaining about the history of the factory and its present production. The teacher pops in and observes the interaction for a couple of minutes and then makes a quiet exit without disrupting the presentation. The guests ask a few questions, and the speaker explains some more. At the end, all four students get up and walk into the school workshop to observe the production process. The other host takes over and explains how the paper machine works. There is quite a lot of noise in the workshop; the speaker almost has to shout. At the end of the tour, the speaker asks if the guests have any more questions and, since they do not, the hosts wish the guests goodbye. The students then fill in self-assessment and peer assessment sheets. The following week's lesson is

spent reflecting on and discussing the simulations and the peer and self-assessments.

This assessment activity helps vocational college students learn factory presentation skills in English. The task is a fairly realistic simulation of one of their possible future tasks in the workplace. The assessment is an integrated part of other learning activities in class, in that the class starts preparing for it together by discussing the properties of a good factory tour, and they use another couple of lessons for planning the tours and practising the presentations. Working in groups makes efficient use of class time, and having students rate themselves and their peers further supports student reflection of what makes a good factory tour. Pair work in preparing the presentation simulates support from colleagues in a workplace. The teacher's main role during the preparation stage is to structure the activities and support the students' work. During the assessment event he circulates among the groups and observes each pair for a couple of minutes, and after the event he evaluates performances, conducts assessment discussions with each pair, and makes notes on their peer and self-assessments for future use in grading.

Scenario 4
The interviewer and the examinee are talking about the examinee's job. The interviewer asks her to compare her present tasks to her earlier job, and then to talk about what she would like to do in the future. And what if she were to move abroad? This is obviously not the first time the examinee is talking about her work in English, she has a very good command of the specialist vocabulary, and while her speaking rate is not very fast, this may be how she speaks her mother tongue, too. She has no problem answering any of the interviewer's questions. In around fifteen minutes, the interviewer winds down the discussion and says goodbye to the examinee. She has made an initial assessment of the performance during the interview, and now she makes a final evaluation and writes it down on an assessment form. Then she has a quick cup of coffee, after which she invites the next examinee into the room.

The test in Scenario 4 is part of a proficiency test battery for adults. **Proficiency tests** are examinations that are not related to particular learning courses but, rather, they are based on an independent definition of language ability. This particular test is intended for adults who want a certificate about their language skills either for themselves or for their employers. Talking about their profession and their future plans is thus a relevant task for the participants. The certificates that the examinees get report a separate score for speaking.

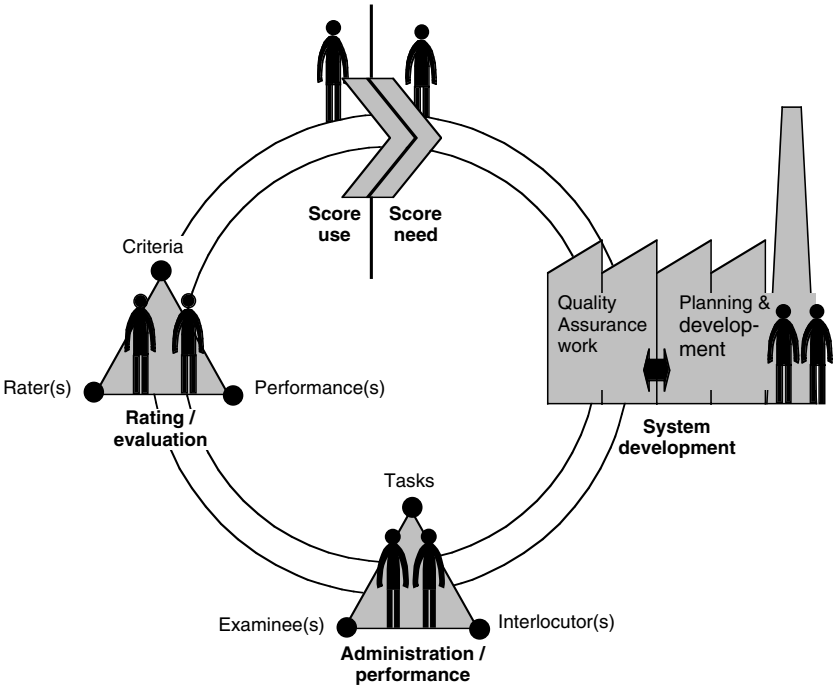
The surface simplicity of the individual interview as a format for testing speaking hides a complex set of design, planning and training that underlies the interaction. This is especially true if the interview is part of a proficiency test, but it is also true in settings where the participants may know each other, such as interview tests conducted by a teacher. This is because, like all tests, the interview should be fair to all participants and give them an equal opportunity to show their skills. Since the test is given individually, the interviewer needs to follow some kind of an outline to make sure that he or she acts the same way with all the examinees. If some of the tests are conducted by a different interviewer, the outline is all the more important. Furthermore, the criteria that are used to evaluate the performances must be planned together with the interview outline to ensure that all performances can be rated fairly according to the criteria. This partly depends on the interlocutor's interviewing skills, and in big testing organisations interviewer training and monitoring are an essential part of the testing activities. The interviewers in the test of Scenario 4 are trained on a two-part workshop and then conduct a number of practice interviews before being certified for their job.

The cycle of assessing speaking

As the examples above show, assessing speaking is a process with many stages. At each stage, people act and interact to produce something for the next stage. While the assessment developers are the key players in the speaking assessment cycle, the examinees, interlocutors, raters and score users also have a role to play in the activities. This book is about the stages in the cycle of assessing speaking and about ways of making them work well. It is meant for teachers and researchers who are interested in reflecting on their speaking assessment practices and developing them further.

A simplified graph of the activity cycle of assessing speaking is shown in Figure 1.1. The activities begin at the top of the figure, when someone realises that there is a need for a speaking assessment. This leads to a planning and development stage during which, in a shorter or longer process, the developers define exactly what it is that needs to be assessed, and then develop, try out and revise tasks, rating criteria and administration procedures that implement this intention. They also set up quality assurance procedures to help them monitor everything that happens in the assessment cycle. The assessment can then begin to be used.

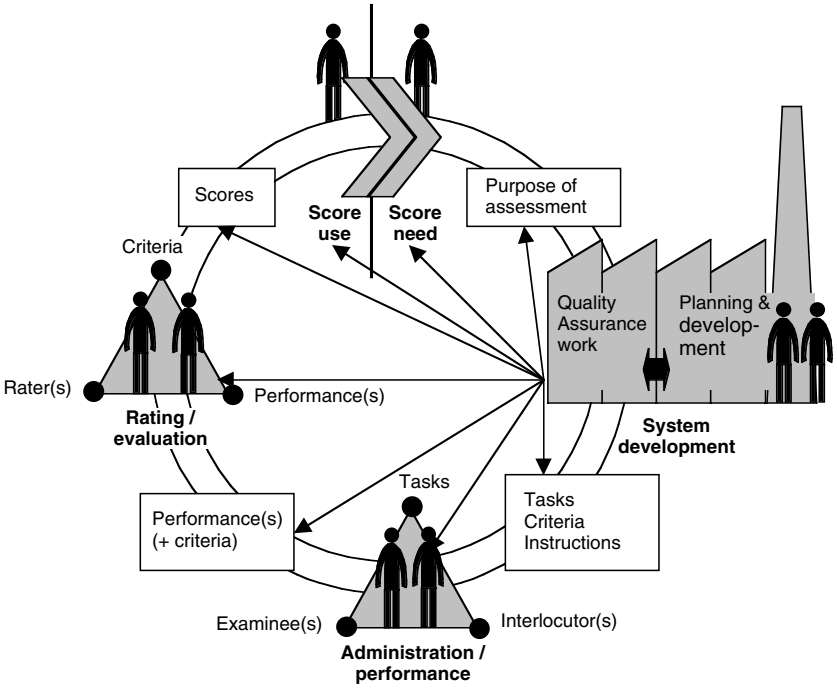
Figure 1.1 The activity cycle of assessing speaking



The cycle continues with two interactive processes that are needed for 'doing' speaking assessment. The first is the test administration/test performance process, where the participants interact with each other and/or with the examiner(s) to show a sample of their speaking skills. This is often recorded on audio- or videotape. The second process is rating/evaluation, where raters apply the rating criteria to the test performances. This produces the scores, which should satisfy the need that was identified when the test development first started. I use the term **score** in a broad sense to refer to numerical scores, verbal feedback, or both. At the end of the cycle, if the need still exists and there is a new group of examinees waiting to be assessed, the cycle can begin again. If information from the previous round indicates some need for revision, this has to be done, but if not the next step is administering a new round of tests.

Figure 1.1 is simplified in many senses, two of which are that, while it shows activity stages, it does not show the products that are taken forward from each stage or the scope of the quality assurance work in the cycle. Before going into these, let me say something about the shapes

Figure 1.2 Stages, activities and products in assessing speaking



used for the stages. At the top of the cycle, score need and score use are indicated by dovetailed arrows. This signifies the need for the start and end of the cycle of assessing speaking to fit together. The second stage is shown as a factory. This is the test developers' workplace. They develop the assessment and produce the documents that are needed (tasks, criteria, instructions) to guide the activities. As at any factory, quality assurance is an important aspect of the development work. It ensures that the testing practices being developed are good enough for the original purpose. Moving along the cycle, the administration and rating processes are shown as triangles because each of them is a three-way interaction. The human figures in the cycle remind us that none of the stages is mechanical; they are all based on the actions of people. Score need and score use bind the stages of assessing speaking into an interactive cycle between people, processes and products.

Figure 1.2 shows the same activity cycle with the most important products – documents, recordings, scores, etc. – and the scope of the quality assurance work drawn in. To begin from the top of the figure, the first doc-

ument to be written after the realisation that speaking scores are needed is a clarification of the purpose of the assessment. This guides all the rest of the activities in the cycle. Moving along, the main products of the planning stage are the tasks, assessment criteria, and instructions to participants, administrators, interlocutors and assessors for putting the assessment into action. At the next stage, the administration of the test produces examinee performances, which are then rated to produce the scores.

As is clearly visible in Figure 1.2, quality assurance work extends over the whole assessment cycle. The main qualities that the developers need to work on are construct validity and reliability. **Construct** is a technical term we use for the thing we are trying to assess. In speaking assessments, the construct refers to the particular kind of speaking that is assessed in the test. Work on **construct validity** means ensuring that the right thing is being assessed, and it is the most important quality in all assessments. Validation work covers the processes and products of all the stages in the speaking assessment cycle. They are evaluated against the definition of the speaking skills that the developers intended to assess. **Reliability** means making sure that the test gives consistent and dependable results. I will discuss this in more detail in Chapter 8.

The organisation of this book

This chapter has given a brief introduction into the world of assessing speaking. The next four chapters deal with existing research that can help the development of speaking assessments. Chapter 2 summarises applied linguistic perspectives on the nature of the speaking skill and considers the implications for assessing the right construct, speaking. Chapter 3 discusses task design and task-related research and practice. Chapter 4 takes up the topic of speaking scales. It introduces concepts related to scales in the light of examples and discusses methods of scale development. Chapter 5 discusses the use of theoretical models as conceptual frameworks that can guide the definition of the construct of speaking for different speaking assessments.

Chapters 6 through 8 then provide practical examples and advice to support speaking test development. Chapter 6 presents the concept of test specifications and discusses three examples. Chapter 7 concentrates on exemplifying different kinds of speaking tasks and discussing their development. Chapter 8 focuses on procedures for ensuring the reliability and validity of speaking assessments. The main themes of the book are

revisited in the course of the discussion. The chapter concludes with a look at future directions in speaking assessment.

In this chapter, I have introduced the activity of assessing speaking. Different assessment procedures for speaking can look very different, they may involve one or more examinees and one or more testers, the rating may be done during the testing or afterwards based on a recording, and the scores may be used for a wide range of purposes. Despite the differences, the development and use of different speaking assessments follow a very similar course, which can be modelled as an activity cycle. The activities begin with the developers defining the purpose of the assessment and the kind of speaking that needs to be assessed, or the test construct. To do this, they need to understand what speaking is like as a skill. This is the topic of the next chapter.

CHAPTER TWO

The nature of speaking

In this chapter, I will present the way speaking is discussed in applied linguistics. I will cover linguistic descriptions of spoken language, speaking as interaction, and speaking as a social and situation-based activity. All these perspectives see speaking as an integral part of people's daily lives. Together, they help assessment developers form a clear understanding of what it means to be able to speak a language and then transfer this understanding to the design of tasks and rating criteria. The more these concrete features of tests are geared towards the special features of speaking, the more certain it is that the results will indicate what they purport to indicate, namely the ability to *speak* a language.

Describing spoken language

What is special about spoken language? What kind of language is used in spoken interaction? What does this imply for the design of speaking assessments?

The sound of speech

When people hear someone speak, they pay attention to what the speaker sounds like almost automatically. On the basis of what they hear, they make some tentative and possibly subconscious judgements about the speaker's personality, attitudes, home region and native/non-native

speaker status. As speakers, consciously or unconsciously, people use their speech to create an image of themselves to others. By using speed and pausing, and variations in pitch, volume and intonation, they also create a texture for their talk that supports and enhances what they are saying. The sound of people's speech is meaningful, and that is why this is important for assessing speaking.

The sound of speech is a thorny issue for language assessment, however. This is first of all because people tend to judge native/non-native speaker status on the basis of pronunciation. This easily leads to the idea that the standard against which learner pronunciation should be judged is the speech of a native speaker. But is the standard justified? And if it is not, how can an alternative standard be defined?

The native speaker standard for foreign language pronunciation is questioned on two main accounts (see e.g. Brown and Yule, 1983: 26–27; Morley, 1991: 498–501). Firstly, in today's world, it is difficult to determine which single standard would suffice as *the* native speaker standard for any language, particularly so for widely used languages. All languages have different regional varieties and often regional standards as well. The standards are valued in different ways in different regions and for different purposes, and this makes it difficult to choose a particular standard for an assessment or to require that learners should try to approximate to one standard only. Secondly, as research into learner language has progressed, it has become clear that, although vast numbers of language learners learn to pronounce in a fully comprehensible and efficient manner, very few learners are capable of achieving a native-like standard in all respects. If native-like speech is made the criterion, most language learners will 'fail' even if they are fully functional in normal communicative situations. Communicative effectiveness, which is based on comprehensibility and probably guided by native speaker standards but defined in terms of realistic learner achievement, is a better standard for learner pronunciation.

There are, furthermore, several social and psychological reasons why many learners may not even *want* to be mistaken for native speakers of a language (see e.g. Leather and James, 1996; Pennington and Richards, 1986): a characteristic accent can be a part of a learner's identity, they may not want to sound pretentious especially in front of their peers, they may want recognition for their ability to have learned the language so well despite their non-native status, and/or they may want a means to convey their non-native status so that if they make any cultural or politeness mistakes, the listeners could give them the benefit of the doubt because of their background.