4 Deciding what to test

1. The test design cycle

In the previous chapters we have considered why testing exists, its origins, what benefits it brings, and what its costs are. We have seen that testing is a socially constructed enterprise that has been part of human civilisation for as long as we can tell. Its function within society can be to maintain the status quo, or it can be the means by which equality of opportunity is afforded to all. Testing is therefore never value free, and is inherently a political activity (Fulcher, 2009). We have shown that there are two main testing paradigms, although we admit that the boundaries between them are fluid. Large-scale standardised testing is the most prevalent form, is usually externally mandated and plays the most important role in selection. Classroom testing, on the other hand, is of most use in the learning process. Each paradigm has its associated technologies, methodologies and challenges.

We now turn to issues of test design. In this and the following chapters we concentrate on the process of test development, and how to evaluate the outcomes. However, we will not be able to leave aside the social, political and ethical issues entirely. They still have an impact upon the key decisions that are made when building a new test or assessment.

We begin with the test design cycle. When asked to produce a test, many teachers start with writing test items. Managers frequently encourage this because they expect a teacher to produce the test in an afternoon, or over the weekend at best. It is something that is perceived to be the easiest part of the role of a teacher. In all fairness to the major testing agencies, they understand that standardised test development takes a long time and is very expensive. It is realised how important tests are for the lives of the test takers, and a great deal of effort is expended to investigate the reliability, validity and impact of the tests concerned (see, for example, Wall and Horák, 2007, 2008; Green, 2007). Writing test content is not normally the starting point of test design, for either standardised tests, or classroom tests.

The test design cycle is illustrated in Figure 4.1. In this chapter we will consider just the first three stages of the cycle, although 'Inferences' and 'Decisions' cannot be separated from our definition of test purpose; other stages will be considered in subsequent chapters.

The starting point in the cycle is normally test purpose. In Chapter 1 we quoted Carroll and Cronbach, warning us that without a clear statement of test purpose there could be no convincing rationale for selecting test content or format. This is equally gest that teachers ask the following questions when considering how to design their assessment for a learning programme:

- What do we hope to achieve?
- How can the important issues be identified?
- Who should be involved?
- What may be the effects of evaluating?
- How can information be collected and analysed?
- Who can best collect and analyse the information?
- What are the most applicable sources of information?
- What are the constraints (e.g. time, manpower, political)?

These are excellent questions that should be addressed. Before deciding to test, there has to be a reason for testing. Defining test purpose also incorporates another critical question: what do we test? Or, what is the information we need, and why do we need it? We define purpose and ask these questions first because it provides the basis of all further decisions. To explain this further, we will consider the arguments put forward by

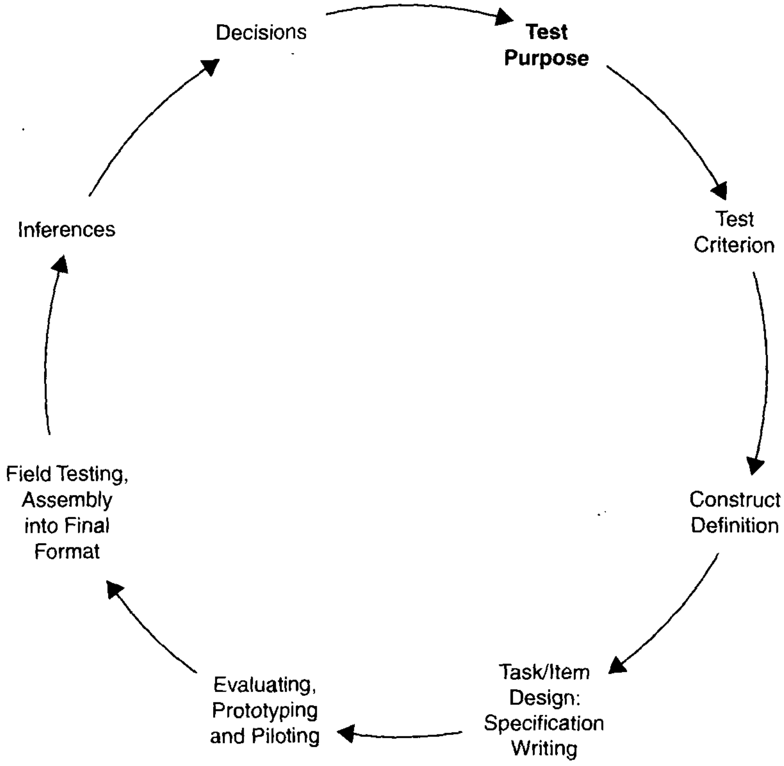
Decisions
Test Purpose
Test Criterion
Construct Definition
Task/Item Design: Specification Writing
Evaluating, Prototyping and Piloting
Field Testing, Assembly into Final Format
Inferences

Fig. 4.1. The test design cycle

Fulcher (2006) and Fulcher and Davidson (2009), who use architecture as a metaphor for test development.

When architects begin to design a building, they must have a very clear idea of its purpose. If a client wishes to open a supermarket there is little point in designing a neoclassical residential town house. If the requirement is for a space to repair cars, the architect would not design a restaurant. And if I wished a to build a country retreat where I could get away from the city, light a fire on a cold winter evening and watch TV, I would be rather upset if the architect produced plans for a block of flats. Similarly, the materials needed for the construction of these buildings would be different, and the cost of the building would vary accordingly.

The same is true of language testing. As Ingram (1968: 70) once refreshingly put it, 'All tests are for a purpose. A test that is made up without a clear idea of what it is for, is no good.' If the purpose of my test is to assess the achievement of the learners in my class on the material covered in the last two months – which is a very common 'linear' requirement of teachers – I would need to design a test that was related to the course. There are a number of ways that teachers can do this. The first would be to sample content directly from the syllabus, and to design assessment tasks that reflect the kinds of processes and skills that were the target of learning. Another might be to look at the learning objectives or outcomes, and to base the assessments on these, rather than directly sampling from the syllabus. Or some combination of these might be used. Short cuts are often provided by course book publishers in the form of a 'test book' to accompany the course. However, teachers should use these with care, as it is not always clear that they provide the kind of learning information that we might need (see Chapter 3).

When tests are used for certification, the need to state the precise purpose of the test is even more acute. If we consider a situation in which it is necessary to certify the reading and writing skills for aircraft engineers the stakes could not be higher. Here there is a clear need to undertake a specified task in the real world that requires the use of language. In this case, the person has to be able to read a technical manual, follow the instructions carefully to inspect an aircraft and repair any faults that are found. At the end of the process they must write a report on what has been done so that it can be signed off by a supervisor to say that the aircraft is fit to fly. If the engineers are not capable of fulfilling these tasks in English, there is a clear and obvious safety hazard.

The purpose of the test is therefore very specific. It is to assess the ability of the test taker to understand the technical manual, to follow instructions provided in the manual, and to write a report in a specified genre that is acceptable to the supervisory engineers. This illustrates the next step on our test development cycle, which is defining the *test criterion*; in this case the criterion is successful use of the manual and effective communication through technical reports in the target domain. In order to study and describe the criterion, the test developer is fortunate in that it is possible to collect a representative sample of manuals that can be analysed. It is possible to design questions based on the sample manuals that can be given to proficient and non-proficient engineers in order to see which task types discriminate well between them (a 'group difference' study). Supervisors can be interviewed in order to discover what kinds of target behaviours are expected, and they can be asked to judge the adequacy of a range of sample reports collected from engineers in order to create a corpus of 'adequate' and 'substandard' reports. The key features of these could be described in order to create definitions of 'masters' and 'non-masters' for the purposes of scoring. In other words, it is test purpose that drives all the other activities associated with the development of a test.

As Fulcher and Davidson (2009: 123–124) put it,

> *a statement of test purpose is likely to include information on the target population and its ability range. Test developers normally state target domains of language use, and the range of knowledge, skills or abilities that underpin the test. This statement justifies the selection of constructs and content by articulating a direct link between intended score meaning and the use to which the scores will be put in decision making.*

Without this level of explicitness, we would have *design chaos*. This is a situation in which we are asked to design a structure without a purpose. Just as it is difficult to evaluate the success of a building without a purpose, it is impossible to evaluate a test. If we have design chaos at the beginning of the process, we have *validity chaos* at the end.

2. Construct definition

We come to the third label in the test design cycle. This is potentially the most difficult to understand and to apply, because the analogy with architecture does not help as much as it does with other aspects of seeing test design as 'building'. Constructs are the abilities of the learner that we believe underlie their test performance, but which we cannot directly observe. These begin as 'concepts', and we can identify them because they are usually abstract nouns. The oldest construct in education is probably 'intelligence', the meaning of which has been argued over for centuries; and it appears that we are no closer to having an agreed definition now than we were 100 years ago (Evans and Waites, 1981). Another construct that teachers often use is 'attitude'; we frequently say that a learner has a 'positive' or 'negative' attitude toward language learning. When judging how well a student communicates, we may talk about their 'fluency'. We can't point to a specific example of 'positive attitude' or of 'fluency'. Our judgement that these things exist, to some degree, is extracted from many examples of things we observe in the behaviour of the individuals concerned.

When we abstract from what we observe and create labels we are essentially building a basic theory to explain observable phenomena. Kerlinger and Lee (2000: 40) define constructs as 'concepts' that are adapted for scientific investigation in two ways. The first is that they are defined in such a way that we can measure them. This is usually termed an 'operational definition', which tells an observer what kinds of 'things' count towards positive attitude. Sometimes these 'things' have to be artificially manipulated in order to get a measurement. Measuring attitude normally has to be undertaken through a survey instrument of some kind, in which a learner is presented with statements regarding language learning and is asked to respond to indicate the degree to which they have a favourable or unfavourable reaction. These responses can be quantified, so that the variable 'attitude' can be plotted on a scale.

The point is that the definition of the construct is something that has to be undertaken carefully if it is to be assessed, as we need to know what it is we have to ask a learner to do, so that we can observe it , and decide whether (and to what extent) this abstract ability is present. Why does this matter? Isn't it good enough just to watch learners and get a general idea of what they can and can't do? Yes, in some circumstances. But we will use a non-language example in order to explain the power of 'construct' language.

Imagine that you are a local government official in a coastal district. In your district there are a number of wonderful beaches where people frequently swim during the warm weather, and for the coming season it is necessary to hire an additional lifeguard to meet health and safety regulations. An advertisement is placed in the local paper, and a fairly large number of people apply. The applicants are shortlisted, and the five finalists are asked to come to your district for an afternoon, at the end of which you will have to make a final choice. The task facing you is to decide what qualities you want your lifeguard to have, and how you will assess the shortlisted candidates against the qualities. What you need is a test (or tests) that can be held in one afternoon. Extending the selection period beyond that is not possible for economic reasons (a major constraint). The purpose of the test is clear: to select a person to work as a lifeguard. The criterion is a little more complex, however. We know what lifeguards are supposed to do: save lives in the sea. But the range of things they actually do is quite diverse. What follows is a summary job description, which is essentially our attempt to describe the criterion – what we expect the person appointed to be able to do when required. What it does not contain is as important as what it does contain, for it creates the frame around the picture that we can work with in designing a test.

The person appointed will be expected to:

- patrol an assigned area of beach, providing surveillance of swimmers, and assistance, including first aid; educate and advise beach-goers of dangerous marine conditions and beach hazards
- respond to emergency incidents, take actions to prevent injury, and perform rescues of injured or incapacitated victims in a dangerous marine environment; administer first aid, CPR and automatic external defibrillation, as necessary
- monitor beach and water population levels; recognise conditions warranting closing or reopening the beach areas
- provide high-level customer service, education and information to residents and beach visitors; identify lost individuals and coordinate reunification
- operate electronic communications, including mobile and portable radios, public address systems and computers.

The person appointed will have the ability to:

- perform physically demanding rescues on the surface and [illegible] and strong currents; make preventive actions under difficult, dangerous and stressful conditions; direct others in emergency situations
- resolve complaints from the public in a sensitive and tactful manner
- communicate clearly and concisely in speech and writing, under stressful conditions
- work for long periods of time in inclement weather conditions.

What are the constructs involved? Potentially there are many, but I will list just four:

Alertness
Affability
Strength
Stamina

If I could give each of the five shortlisted candidates a score on each of these constructs, I would be going some way towards making an informed selection. Of these, we will only discuss stamina. Although we have a general idea of what we mean by stamina, for the particular purposes of this test we define it as 'the ability to maintain arduous physical activity for an extended period of time without fatigue'. The definition of 'extended period of time' can be established by looking at the typical length of time it takes for lifeguards to retrieve swimmers in trouble from the sea. We will assume that this is up to, but rarely exceeding, twenty minutes. The reason this construct and its definition are so important is because we cannot possibly test each candidate under all the possible conditions that they may encounter in real life. If we consider just some of the variables involved, we might have to account for wave height, strength of current, water and air temperature, wind, distance to swimmer, condition of swimmer (unconscious, panicking, underwater, and so on), weight and size of swimmer. The list could go on. However, in a test we simply cannot replicate real life. Bachman (1990: 301–323) discusses what he calls the 'real-life' approach and the 'interactive language use' approach to defining test authenticity. In the real-life approach, we judge the 'authenticity' of the test on the basis of how well it replicates real life in the tasks. But we have seen that there are so many *performance conditions* for our stamina test that they cannot all be replicated. Further, in a test we frequently cannot replicate real life, anyway. In the stamina test we could not ask the candidates to take the test in the sea; health and safety legislation would prevent this. In addition, in the sea we couldn't replicate any particular performance condition that would be the same for all test takers, thus making the test fair. The test conditions would vary from administration to administration, because we cannot control the sea or the weather.

The definition of 'authenticity' that Bachman prefers is that of Widdowson (1978: 80), as 'a characteristic of the relationship between the passage and the reader and it has to do with appropriate response', or 'a function of the *interaction* between the test taker and the test task' (Bachman, 1990: 117, italics in the original). This is where the construct comes into play. The task is designed in such a way that it is an index of the level

an ability that is used in all the performance
might happen in real life. The ability is therefore separate from the instances in which it might be displayed in real life.

Consider the following stamina test (Northwest Lifeguards Certification Test, 2008):

100 Yard Medley – 12 points possible

**Note: A person may not advance their position by using the bottom of the pool or the side walls. The end walls may be used for advancement.

A. Passive Drowning Victim Rear Rescue, 0, 1, 2, 3, or 4 points

1. Approach the victim from behind with the tube.
2. Reach under the victim's armpits and grasp his/her shoulders.
3. Squeeze the rescue tube between your chest and the victim's back.
4. Keep your head to one side of the victim's head to avoid being hit.
5. Roll the victim over so that they are on top of the rescue tube.
6. Move the victim to the other end of the pool.

B. Cross Chest Carry for 25 yards, 0, 1, 2, or 3 points

1. If the rescuer chooses to use the scissors kick, the hip must be in the victim's back; doing the whip kick, the rescuer must be on their back.
2. Hands must grasp behind the armpit.
3. Victim must be secure and controlled.
4. Victim must be level with face clear of the water.

C. Single Armpit Assist on Back for 25 yards, 0, 1, or 2 points

1. Thumb must be up on the inside of the armpit.
2. Eye contact must be maintained (except for quick glances forward for direction).
3. Rescuer must be PUSHING the victim with a smooth motion, no jerking.

D. Underwater Swim for 25 yards, 0, 1, 2, or 3 points

1. The rescuer has 3 seconds after placing the victim on the wall to begin the underwater swim. After the first 3 seconds, each 3 seconds after that will count as 1 breath (see D-3).
2. Rescuer may use the wall to push off underwater, but not over the top of the water. The rescuer must submerge vertically down and push off underwater from the wall.
3. Rescuer may come up once for air without losing points. More than once, a point is deducted for each time.
4. When the rescuer comes up for air, 3 seconds are allowed to get a breath and go back down.
5. Any part of the body breaking the surface is counted as coming up for air.
6. A person may not advance their position in the water (stroke at surface), when coming up for air.

A SCORE OF 0 POINTS ON SECTION D WILL AUTOMATICALLY FAIL A CANDIDATE.

The first thing to note about this task is that it is very controlled. The victim to be rescued is a real person, but the guidelines (not reproduced here) state the size and weight restrictions. Further, they do not struggle, but remain as limp as possible during the test. The test takes place in a pool, and practices that are not allowed (because they would not be available in 'real life') are clearly stated. Most importantly, the criteria for successful completion of each stage of the test are explicit.

The validity question that we face is: how good is the argument that performance on these tasks, as reflected in the scores, is an index of the construct 'stamina'? And, further, to what extent is our *measurement of 'stamina' a predictor* of successful performance in the real world?

To return to language testing, we can apply the same thinking to the reading test that we might construct for our aircraft engineers. Although this is a fairly limited language use domain, we recognise that we cannot describe everything that they might be expected to do. Nor, as Clark (1975: 23) pointed out, is it possible to place test takers into the actual role of an aircraft engineer and 'follow that individual surreptitiously over an extended period of time' to see if they carry out the tasks successfully in English. Not only would it be too costly, it would be unethical, and potentially extremely dangerous. The test must therefore be *indirect*, a number of steps removed from reality. In this sense, it is impossible to have a *direct test*, if this term is interpreted as actually doing what people do in real life as it was in the communicative language testing movement (Morrow, 1979). Designing a test for the engineers is not therefore simply a matter of conducting a job analysis and sampling for the test (Fulcher, 1999), although the analysis will be a crucial element of the research that will inform the selection of relevant constructs that will subsequently drive test design and inform the selection of texts for inclusion on the test.

What constructs are likely to be important for the engineers to undertake their tasks successfully and safely? Grabe and Stoller (2002: 21–30) describe reading constructs in terms of lower and higher processes. Among the lower processes, it is recognised that one of the most important constructs in reading any text is lexical access, defined as the speed with which readers can call up the meaning of a word when it is encountered on the page. This must be automatic if the reader is not to slow down to an extent that the general meaning of a text is lost because of short-term memory overload. Of equal importance is the syntactic parsing, or the ability to recognise groups of words in order to understand word ordering, subordination and other relationships between clauses. If lexical access and parsing are automatic, readers are able to access 'semantic proposition information', which allows clause-level meaning. Among the higher level constructs are the formation of a 'mental text model', which represents in the mind the main propositions of a text together with its supporting ideas and examples. The whole process may be supported or hindered by the activation of relevant background knowledge, as well as familiarity with the discourse organisation of the kinds of texts involved. A general description of reading constructs can be listed in more detail if necessary (Grabe, 1999), but for our purposes this will do. If the most important task for the engineers is to understand a process as it is described and follow that process exactly when checking and repairing an aircraft component, which, if any, of these constructs are relevant?

The first construct that is relevant to the engineers is lexical access to the technical vocabulary that they require in English. The second important construct is the ability to recognise and understand clausal relationships that indicate sequence and action. In order to make sense of information across clauses in texts that describe processes, it is essential to be able to maintain track of cohesive devices, which is a procedural skill to maintain a frame of reference (Fulcher, 1998b). The clausal relations that are most likely to be important in this context are situation–problem–response–evaluation (Hoey, 1983, 62–80), which can be illustrated in the following example:

> *(1) The release of oxygen masks should be checked at the required intervals only when the aircraft is outside the hangar. Personal cleanliness is imperative. Wash dirt, oil and grease from hands before working with equipment. Do not service during fuelling operations as oxygen under pressure and petroleum products may ignite when brought into contact. Release oxygen masks by activating the central release panel and ensuring that each mask in the aircraft has been released and a flow of oxygen is present. (2) The most frequent cause of non-release is the tube becoming twisted or catching on the housing. (3) Examine tangled tubes, gently releasing the mask and replacing as shown in diagram 3b. Any damaged tubes should be replaced with component 284XOM (4) to ensure smooth and easy release of the mask.*

(1) Situation
Containing directions/instructions ('Wash … Do not service … Release … ')
(2) Problem
(3) Response
(4) Evaluation

Examples of clausal relations can be analysed from manuals so that suitable task types can be constructed. Along with a test of cohesion, the development of a test of technical vocabulary might be helped by the creation of a corpus of maintenance manuals to produce essential vocabulary lists. The ability to recognise steps in processes might be assessed using visuals, as might the ability to select the correct action for a given situation.

To define constructs of interest, language tests need to be developed for clearly specified purposes. A general language test would not target the precise constructs that underlie performance in any particular domain. This is one of the major problems with using 'off-the-peg' solutions for readiness to use language in specific domains. The scores on 'general' language tests are not necessarily built on constructs relevant to the decisions that need to be made in a specific context. If test scores vary because of constructs or other test features that are irrelevant to the decision we wish to make, it is construct irrelevant variance.

Constructs therefore need to be selected for their applicability to the context of test use. When we do this, we place the description of how we made these decisions in a document that narrates the design process. This document links the constructs to the purpose and context of the test, and is part of the *test framework* (Chalhoub-Deville, 1997). It forms part of the evidence to show that any inferences we make about the meaning of test scores are valid, and can be presented as part of the validation evidence.

3. Where do constructs come from?

Fulcher and Davidson (2009: 126–127) describe constructs as the 'design patterns' in architecture, which can be selected for use in a particular building project. They are abstractions with operational descriptions that can be relevant to many contexts in different ways, but are not useful in all contexts. The ability to comprehend processes and carry out procedures is highly relevant to aircraft engineers. It is also relevant to students who may be carrying out experiments in a laboratory. The context and materials are different, but the constructs may be largely the same. But these constructs will not be relevant to a test for a language test for tour guides, for instance. They may be partially relevant to a test for a travel assistant who may have to follow certain procedures in booking vacations for clients, but this would be a very small element in the broader communicative requirements of the context. If such a test only contained procedural questions, it would be *construct under-representative.* This means that the constructs tested are not enough to ensure that the score represents the abilities that a person may need to succeed in this role.

Constructs are usually described in *models.* These operate at a higher level than the test frameworks. They are general and abstract by nature, whereas the test framework is a selection of constructs that are relevant to a particular purpose. The relationship between models and frameworks is illustrated in Figure 4.2 (from Fulcher and Davidson, 2009: 127). You will notice that underneath a framework is a test specification. We will discuss test specifications in Chapter 5. Together, these three levels constitute the levels of test architecture, from the most abstract (models) to the most specific (specifications).

This conceptualisation recognises that there are a very large number of language constructs and infinite language use situations, but that only some constructs or abilities will be realised in each situation. Further, even the most proficient language users will not be equally 'proficient' in all contexts. As Lado (1961: 26) put it, 'The situations in which language is the medium of communication are potentially almost infinite. No one, not even the most learned, can speak and understand his native language in any and all the situations in which it can be used.' Lado was ahead of his time. He realised that language was not only a tool for humans to get things done, but the means by which personal and cultural meaning was encoded.

> *Language is more than the apparently simple stream of sound that flows from the tongue of the native speaker; it is more than the native speaker thinks it is. It is a complex system of communication with various levels of complexity involving intricate selection and ordering of meanings, sounds, and larger units and arrangements.*
> (1961: 2)

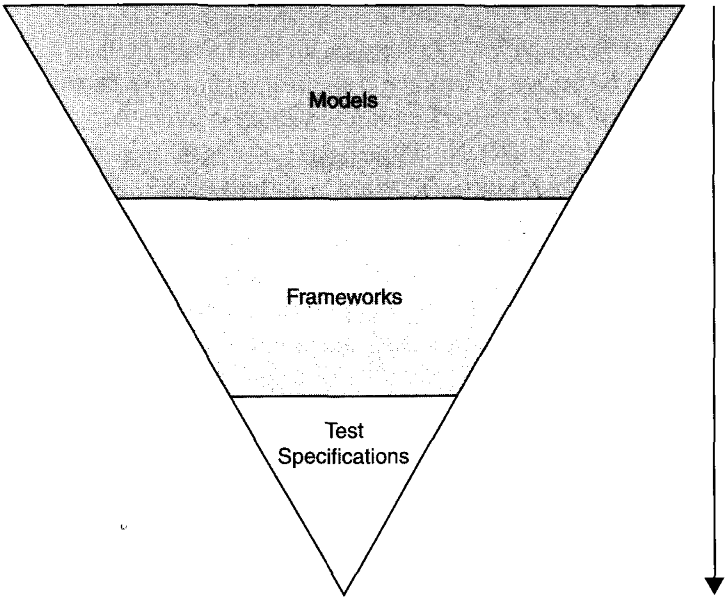

Fig. 4.2. The levels of architectural documentation

Lado (1961: 6) was the first to create a model of language use that contained constructs, as shown in Figure 4.3. For communication to be successful, he argued that language use had to be automatic – or 'habit' – as he put it in the behaviouristic terminology of the time. Lado's understanding is the same as that held in current SLA research: 'automaticity refers to the absence of attentional control in the execution of a cognitive activity' (Segalowitz and Hulstijn, 2005: 371). This automatic processing is important for the language user to select the form (sounds, words, grammatical units) to create meaning. The distribution is the 'permitted environments in which each form appears' (Lado, 1961: 5). This would include concepts such as collocation and colligation.

The linguistic meanings can only be understood if there is also an understanding of the cultural meanings behind the encodings. Lado said that it was difficult to describe these culturally bound meanings, but he saw them as the ways in which people from certain backgrounds used language to communicate the 'organization of behaviour', of ways of doing and of believing, through form, meanings and distributions. In order to penetrate meaning, cross-cultural communication issues were seen as critical, and could even be tested. Yet, Lado approaches this topic of differences between cultures with 'a basic assumption of and belief in the unity of all mankind. All races have the same origin and are capable of the same emotions and the same needs encompassing the whole range of human experience from hunger and the craving for food to theological inquiry and the seeking of God' (1961: 276). In this inspiring writing, Lado goes on to say that, apart from the culturally embedded meaning, individuals bring their own personal meaning which comes from life experiences. This stands outside culture, and

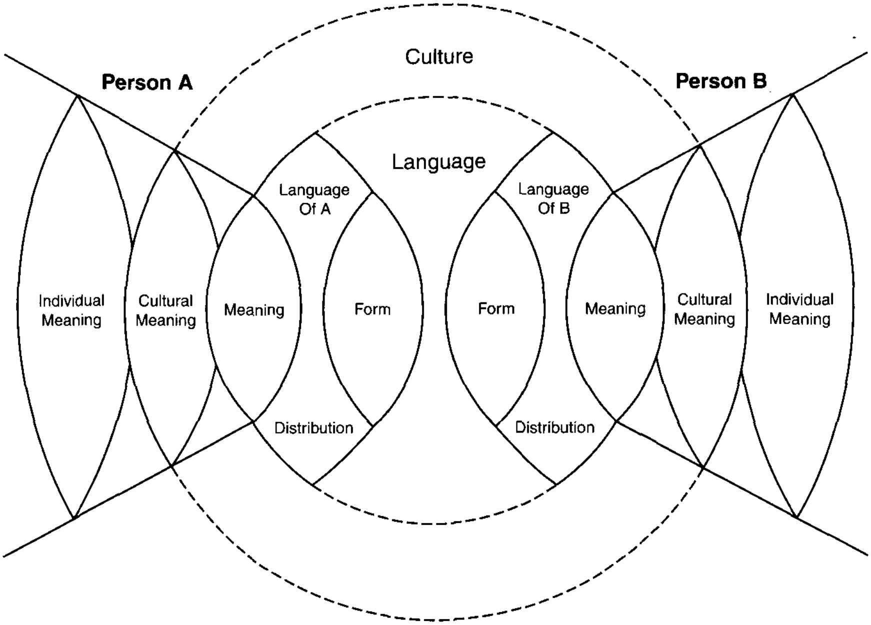

Fig. 4.3. Language, culture and the individual

represents the individuality that is expressed through language as an act of being. This model exists within each of the four skills of reading, writing, speaking and listening, and their interactive use (listening and speaking), in an infinite number of situations.

It is difficult to image now just how revolutionary Lado's argument was in 1961. At the time there was an assumption that language tests should limit themselves to language alone, usually considered to be grammar, lexis and the sound system. Carroll (1958: 8) is a typical example of this position, when he argues that 'it is assumed that we are concerned solely with the acquisition of a foreign language, not with the acquisition of the culture of a foreign people, nor the appreciation of its literature'.

Lado's work represents the beginning of model development for language and communication. This model, like others, can serve as an heuristic that can help us to select the constructs that are relevant to a test or assessment we need to design for a particular context. What it cannot do is tell us which forms we may wish to test, or what types of intercultural communication we should focus on, for any given context. That is the role of the test designer when creating a *test framework*. For example, if I wished to construct a test for shopkeepers in a tourist district, one element of intercultural communication that I may wish to test is the ability to establish rapport. In order to do this, it is critical to have the linguistic resources to follow a sales script, while also appearing friendly. It has been discovered that failure to comply with the cultural expectations of what happens in sales encounters can lead to hostility and loss of business (Ryoo, H.-K., 2005). We discuss this particular example in greater detail in Chapter 7.

Moving backwards and forwards between the architectural levels of model and framework is a challenging and fruitful way to look simultaneously at what happens in communicative situations and the constructs required for successful communication. It is also possible to use multiple models to inform choices. Other work on intercultural communication (Byram, 2000: 9–10) suggests that the following five additional abilities underlie the successful understanding of unfamiliar cultural meaning:

- attitudes: curiosity and openness, readiness to suspend disbelief about other cultures and belief about one's own
- knowledge: of social groups and their products and practices in one's own and in one's interlocutor's country, and of the general processes of societal and individual interaction
- skills of interpreting and relating: ability to interpret a document or event from another culture, to explain it and relate it to documents from one's own
- skills of discovery and interaction: ability to acquire new knowledge of a culture and cultural practices and the ability to operate knowledge, attitudes and skills under the constraints of real-time communication and interaction
- critical cultural awareness/political education: an ability to evaluate critically and on the basis of explicit criteria perspectives, practices and products in one's own and other cultures and countries.

Some, but not all, of these, may inform the development of a test or a related language course.

4. Models of communicative competence

Since Lado, models have evolved as we have learned more about what it means to know and use a language. They have also taken a variety of forms. Perhaps the most common models are those that attempt to present an abstract overview. The grain size of these models is very large. There are other models that attempt to reduce the grain size and give a great deal of detail about what kinds of competences or abilities are relevant to particular kinds of interaction. There are problems with both types. The more general and abstract models are much more difficult to apply to a particular context when we write our test framework. The more detailed models, on the other hand, often attempt to describe contexts of language use in ways that we do not recognise as being descriptions of what actually happens in communication. They lack both a theoretical and empirical basis for the detail they purport to offer.

Whatever their grain size, all models have been deeply influenced by the work of Hymes (1972). Breaking away from traditional linguistics, he argued that humans have an ability for language use that fulfils a social function. This ability, he argued, can be defined in terms of four kinds of knowledge. First of all, we know whether it is possible to say something (using the grammatical and lexical resources of the language).

Secondly, we recognise whether it is feasible to say it, even if it is grammatically possible. Thirdly, we know whether it is appropriate to the context. This kind of knowledge can be acquired through social interaction. And fourthly, we know whether or not something actually occurs (or is said), even if it is possible, feasible and appropriate.

In what is probably the most quoted section of any paper in applied linguistics, Hymes (1972: 15) says:

> *Attention to the social dimension is thus not restricted to occasions on which social factors seem to interfere with or restrict the grammatical. The engagement of language in social life has a positive, productive aspect. There are rules of use without which the rules of grammar would be useless. Just as rules of syntax can control aspects of phonology, and just as semantic rules perhaps control aspects of syntax, so rules of speech acts enter as a controlling factor for linguistic form as a whole.*

Hymes therefore talked about 'competence for grammar', and 'competence for use', as two aspects of knowledge. This is contrasted with performance – the actual use of language – which is judged by its 'acceptability' in the context of use (1975: 18). As Shohamy (1996: 139) notes, this early model brings into the picture non-linguistic elements to models of competence that have influenced all further developments. This is not to say that Lado's model did not have non-linguistic elements; it certainly did, but it did not have the notion of an underlying 'ability for performance' that later models contain.

In the rest of this discussion we will attempt to critically describe various models, beginning with mainstream models that attempt to define constructs underlying performance. We will then turn to performance models that try to describe language in behavioural or functional terms, rather than in terms of theoretical constructs.

Construct models

Perhaps the most influential post-Hymes model is that of Canale and Swain (1980). They draw directly on Hymes to create a model that could be used as the basis for syllabus or test design for communicative purposes. It is interesting to note that Canale and Swain's work began with a consideration of teaching and syllabus design rather than testing. It was not assumed that testing would drive teaching. Rather, they thought that a consideration of what it means to 'know a language' for use in context would inform both. Placing this understanding before the design of a teaching syllabus or a test had in fact long been part of sound planning (Spolsky, 1968). Communicative competence, they claimed, consisted of three components. The first was grammatical competence, which included knowledge of grammar, lexis, morphology, syntax, semantics and phonology. The second component was sociolinguistic knowledge, or the rules or use and discourse. Finally, they included strategic competence, which they defined as the ability to overcome communicative difficulties. Communicative competence was separated from 'actual performance' in real-life contexts. They argued that a theory of performance as advocated by Hymes was impossible, as it would include all the non-linguistic variables that could affect communication, such as affective factors. Subsequent models have reintroduced non-linguistic variables. Lado (1961: 290–298) recognised that many non-linguistic variables would indeed impact on language learning, and performance on language tests. He included factors such as educational background, insight into language and culture (both one's own and those of others), attitudes towards minority groups, and interest in other peoples. But Lado knew that these things were too complex to take into account completely, or test with any reliability.

Nevertheless, the Canale and Swain model was soon expanded. In two subsequent papers, Canale (1983a, 1983b) began to include performance in the model, under the term 'actual communication', to mean: 'the realization of such knowledge and skill under limiting psychological and environmental conditions such as memory and perceptual constraints, fatigue, nervousness, distractions and interfering background noises' (Canale, 1983a: 5). The expanded model suddenly became much more complex; it included for the first time not only linguistic knowledge, but psychological and contextual variables that would need to be modelled in the design of tests. Sociolinguistic competence was expanded to include pragmatics, including non-verbal behaviour and awareness of physical distance in communication, and discourse competence became a separate category, incorporating knowledge of textual organisation, genres, cohesion and coherence. This expanded model is portrayed in Figure 4.4.

The next adaptation of this model was undertaken by Bachman (1990), who altered it in two important ways. Firstly, he more clearly differentiated between what is classified as 'knowledge' and what is a 'skill'. Secondly, he attempted to show how the various elements of the model interacted in language use situations. To this end, Bachman separated out strategic competence, which is said to include all communication strategies, rather than just compensatory strategies, from two separate knowledge components. The first knowledge component is language competence, and the second is knowledge

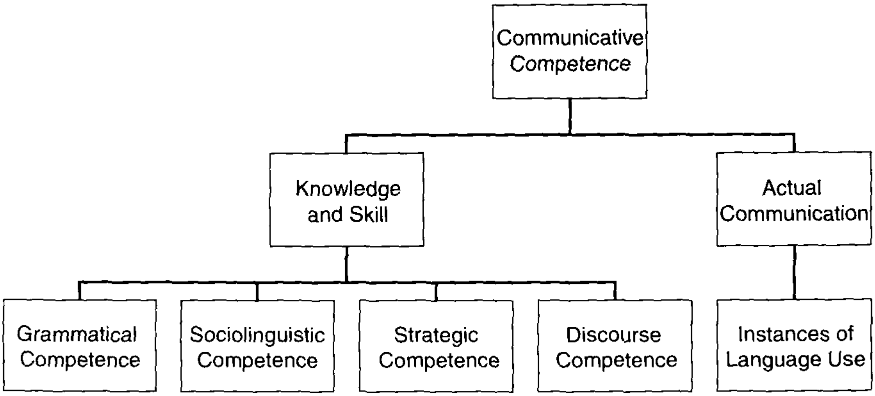

Fig. 4.4. Canale's expanded model of communicative competence

of the world. Both of these are said to affect strategic competence, and how we use it to communicate. In other words, it is strategic competence that draws on knowledge from both linguistic and non-linguistic competences to make communication possible. Strategic competence itself is said to consist of three separate components:

The assessment component:

- identifies the information we need to achieve a communicative goal in a specific context
- decides which language competences are needed to achieve the goal
- decides what knowledge and abilities we share with our interlocutor
- evaluates the extent to which the communication is successful.

The planning component:

- gets information from language competence
- selects modality and/or channel
- assembles the utterance or output.

The execution component:

- makes use of psychophysical mechanisms to realise the utterance.

While world knowledge encompasses what Lado would have called cultural and personal knowledge, language competence is more carefully described in Bachman, as shown in Figure 4.5.

The left-hand side of the Bachman tree contains the traditional linguistic components, while the right-hand side of the tree, under the new title 'pragmatic competence' lists the knowledge necessary to produce appropriate language. Illocutionary competence requires some further explanation. It draws on speech act theory (Austin, 1962), but speech acts are presented in terms of Halliday's (1973) language functions. Ideational functions are concerned with expressing propositions, information and feelings. Manipulative functions are concerned with affecting the world around us, or getting things done, including managing relationships with others. Heuristic functions are related to extending our knowledge of the world through questioning and learning, while imaginative functions concern using language for humour or aesthetic purposes.

Another element of communicative competence that has not appeared in the models we have considered so far is interactional competence. This was first proposed by Kramsch (1986), and it is a competence that relates to any use of language that involves real-time communication with others. The primary observation to support this construct is that, when individuals take part in spoken interaction, the 'text' – what is said – is co-constructed by the participants in the talk. No individual is completely responsible for any part of the talk, not even their own contribution, for it depends upon what else has been said, by whom and for what purpose. In speaking tests, for example, we have long known that proficient interlocutors are prepared to support, or 'scaffold', the speech of the test taker (Ross and Berwick, 1992). More recently, evidence has shown that the discourse style of the interlocutor can affect the performance of the same indi-

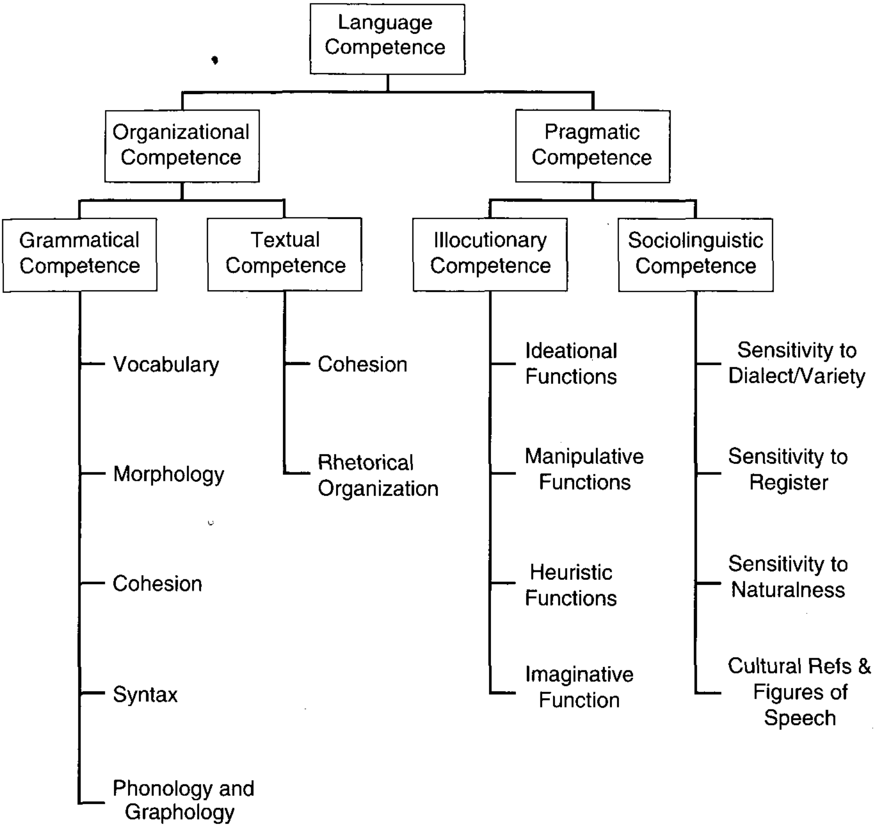

Fig. 4.5. Bachman's components of language competence (Bachman, 1990: 87)

vidual to such an extent that they may get a different score with a different interlocutor (Brown, 2003).

The nature of this competence is highly problematic, however. He and Young (1998) have argued that, because discourse is co-constructed between individuals, it does not make sense to say that interactional competence is something that an individual 'has', in the same way that we could say a learner has a certain degree of grammatical competence, for example. Nor, it is claimed, is this competence 'independent of the interactive practice in which it is (or is not) constituted' (He and Young, 1998: 7). The problem here is that there is no 'home', so to speak, for interactional competence. Young (2008: 101) makes this claim even more strongly:

> *Interactional competence is a relationship between the participants' employment of linguistic and interactional resources and the contexts in which they are employed ... Interactional competence ... is not the ability of an individual to employ those resources in any and every social interaction; rather, interactional competence is how those resources are employed mutually and reciprocally by all participants in*

> *a particular discursive practice. This means that interactional competence is not the knowledge or the possession of an individual person, but it is co-constructed by all participants in a discursive practice, and interactional competence varies with the practice and with the participants.*

This understanding leads to difficulty in using interactional competence as a construct in practical language testing. As Weir (2005: 153) puts it, 'The real problem is that an individual's performance is clearly affected by the way the discourse is co-constructed with the person they are interacting with. How to factor this into or out of assessment criteria is yet to be established in a satisfactory manner.'

However, I believe that accepting the basic premise of interactional competence being disembodied from the individuals taking part in an interaction is flawed. Sociocultural explanations of constructs confuse the ability of an individual to recognise the contextual constraints and freedoms in communicating using available resources, with the actual communication itself. No one would wish to deny that each instance of communication, each instance of new discourse, arises within a context and is dependent upon the variables at play. But this does not mean that the competence to engage successfully in these new interactions is also newly generated in each instance of communication.

We can, however, understand why some applied linguists wish to avoid the distinction between individual interactional competence and its realisation in performance, which leads them to view interactional competence as a disembodied context-dependent phenomenon. It is the rejection of the neo-platonic distinction between competence and performance in the work of theoretical linguists like Chomsky (1965: 4). Recent trends in applied linguistics have favoured social rather than cognitive or psycholinguistic theory and research (Lantolf, 2000, 2002; Lantolf and Poehner, 2008b; Lantolf and Thorne, 2006), and this has also impacted upon language testing (McNamara, 2001; McNamara and Roever, 2006). With this approach, as we saw in Chapter 3 in our consideration of Dynamic Assessment, meaning resides in context and interaction, rather than individuals. It is therefore inevitable that sociocultural interpretations of competence understand all meaning as entirely local and non-generalisable. With every small change in every contextual variable, interactional competence changes. Each interaction is interpreted as an entirely unique realisation of the construct. When second language acquisition researchers have adopted this position, they have normally abandoned the use of the term 'competence' in favour of the word 'capacity' (Fulcher, 1995) to indicate that there is nothing inherently stable about an individual's ability for language use. This is a fundamentally postmodern understanding of language use. It reduces meaning to fleeting social interaction; it takes away from individuals their personal coherence as language-using human beings expressing their own being and existence. Indeed, sociocultural theory even contends that being and identity are only co-constructed through social interaction in which language is used, and that identity 'refers to what we do in a particular context, and of course we do different things in different contexts' (Young, 2008: 108). Nevertheless, Young is aware of the first (although less important) sense of identity as 'self-hood attached to a physical body' that remains the same over time, even though it changes and develops. But this is the least important meaning for sociocultural theory; it is merely an uncomfortable reminder that to have interaction it is necessary to have individuals.

My own view is that the stable individual takes priority. Without this emphasis on the willing participation of individuals in interaction there is no sense in which we hold any power over our own lives and our own sense of identity, which we express through language. The context does not control my contributions to an interaction, other than in the sense that I am sensitive to appropriacy (in Hymes' terms). But this is because I have acquired the necessary pragmatic competence through experience. Nor does the context define who I am. Nevertheless, the insight of sociocultural theory is not to be denied. Precisely how I interact with others in any context is dependent upon how they choose to interact with me.

> *The distinction between what is inside (knowledge, competence, capacity etc.) and what is external (interaction, communication, co-construction etc.) is fluid. The two interact. My 'competence' in Chinese is non-existent; I cannot co-construct discourse and meaning in Chinese. My competence in Greek is fair. I have a 'knowledge' of the vocabulary and structure of the language. Through interaction with other speakers I have gained other competencies which are useful in new situations, but not all. The strategies that I use to interact are simultaneously internal and external to myself. I am recognisable as myself when I speak in a variety of contexts, and yet in context my speech is always contextually bound.*
> (Fulcher, 2003a: 20)

Young (2008: 101–102) provides a simple example of co-constructed discourse between two teachers passing each other on a school corridor:

Ms Allen: How are you?
Mr Bunch: Fine.
Ms Allen: That's good.

Young's analysis of this exchange is that the participants understand the process of turn-taking, that they can recognise when one turns ends and another one may start (transitional relevance places), and that questions and answers (with a response) are three-part conversational structures. In recognising the exchange as a formulaic greeting, each understands that this is part of a particular 'discursive practice'. Young (2008: 102) argues: 'The practice is co-constructed by both participants and in their skilful co-construction they are displaying interactional competence. Interactional competence arises from the interaction and is based on knowledge of procedure and practice that is shared by the participants.' He admits that this interaction is 'based on knowledge of procedure and practice' within the individual, but does not wish to call it 'interactional competence'. This may be illusory. Young himself asks what would have happened if Mr Bunch had responded something along the lines of: 'Well, I'm not too good actually. I had a meeting with my doctor yesterday and he said that I've got a stomach problem that may be stress related. In fact, I have had quite a bit of trouble with my wife recently, she's been ...' Young says that this would be changing the discursive practice in a context that does not lend itself to such interaction. I would certainly agree with this. But the reason why the two participants can engage in formulaic greetings is because the participants possess interactional competence: they know what the rules of interaction are in this particular context, using this kind of formulaic language. Using Hymes' fourth kind of knowledge, they know that extended personal exchanges do not occur in such settings. Mr Bunch's interactional competence is the reason that he does not talk about his health.

It can therefore be argued that interactional competence is something much more substantial than a fleeting social co-construction. It is the knowledge of what is possible and what can happen in certain contexts, applied successfully to managing an interaction, with associated adaptivity to the speaker and other contextual variables.

We can see that there are two approaches to describing interactional competence. One approach sees it as a temporal, non-generalisable, social phenomenon. The other sees it as a set of abilities that an individual brings to the temporally bounded interaction set within a specific social context and discursive practice. If it is conceptualised in the former way, there is little that language testers can do with the construct, because testing is primarily about assigning a score to an individual that adequately reflects the ability of that individual. A group score may be useful in a classroom setting where decisions are made regarding the next task to be set to the group. In high-stakes testing, however, a group score is neither practical nor useful. This is why the construct of interactional competence has impacted on test method, rather than test content, or the way in which test performance is scored. The introduction of pair- or group-speaking methods in some large-scale tests is seen as a way of tapping into a richer interactional construct. The main claim is that when learners interact with each other it reduces the kind of 'interview language' that characterises interactions with a more powerful examiner-interlocutor (van Lier, 1989; Johnson, 2001).

Studies that have investigated the use of this format, with a focus on interaction between participants, have produced varied results. Many show that the percentage of speaking time for learners increases, as do the number and variety of language functions used. However, reservations have also been expressed. It has been suggested that test takers who are paired with partners of different ability levels, different nationalities, with whom they are unfamiliar rather than familiar, or even extroverts (if one is an introvert), may be at a disadvantage (see Fulcher, 2003a: 186–190, for a summary of the research). Studies conducted within a sociocultural paradigm have shown that, all other things being equal, learners in pair or group formats tend to produce more varied language and score slightly higher than they would otherwise on the same task (Brooks, 2009). However, these studies have nothing at all to say about how interactional competence might be scored. Indeed, it is noticeable that the rating scales used in these studies rely on the more established constructs of accuracy, fluency, ability to communicate, coherence and complexity of language. A study by May (2009) recognises that the problem lies both defining and operationalising the construct, but approaches this through rater perceptions rather than identifying the abilities that successful interactants bring to the interaction. She discovered that it is difficult for raters to identify the interactional abilities of individuals without a rating scale that guides the process, and also draws attention to the problems raters have in assigning any score at all to candidates who are less dominant in asymmetric discourse.

The story of research into pair- and group-speaking tests shows that construct definition within models is not entirely theoretical. It is an essentially practical activity. As teachers, we are constantly trying to describe abstractions from observation, and then relate them back to precisely what kind of observations gave them their genesis. In this to and fro between the intensely practical matter of whether we test speaking in groups, pairs and with individuals, we raise questions that pertain to teaching and learning, as well as to assessment: just what counts as evidence for 'interactional competence'? Unless we can produce an operational definition, it cannot be assessed, or taught.

Performance models

Rather than engaging with theoretical notions of competence, many language testers have relied on purely operational definitions of what it means to know and use a language. This began with the communicative language testing movement of the late 1970s and 1980s. There was a reaction against theory, against the use of multiple-choice items, and against what some called the 'statistical sausage machine' that treated human beings as 'subjects' (Underhill, 1987: 105). This was seen as a revolution in which 'there is more blood to be spilt yet' (Morrow, 1979: 156). Communicative language testing was more concerned with tasks and with test content. The closer these resembled what happened in teaching, the better they were. This has become known as an appeal to face validity: what looks good to an experienced teacher probably is; and there isn't much else to be said about testing. The evaluation of performance in a communicative language test was done purely on the basis of behavioural outcomes, defined as the degree to which the test taker achieved the intended communicative effect (see Fulcher, 2000a, 2010, for an extended discussion).

As Shohamy (1996: 145) correctly observes, 'The result was "theory-free" language tests, mostly performance-based, task driven, and considered to be communicative, functional, authentic, and direct.' The immediate appeal of this approach was that it was fairly easy to understand; it appealed to both teachers and the public who did not want to deal with the complexities of language competence and its realisations. But as we have seen in our discussion of the lifeguard test, the appeal to 'directness' and 'authenticity' is illusory. We are rarely concerned with single performances, bound as they are with all the contextual variables that we have outlined. Shohamy (1996: 147) is therefore correct when she says rather bluntly that 'the current performance-communicative task-oriented approach is wrong, simplistic and narrow'.

Nevertheless, performance models have flourished. The most widely used of these is the Common European Framework of Reference (CEFR) (Council of Europe, 2001). We will use this as our primary example to illustrate the nature of performance models.

The CEFR originated in the communicative language teaching movement of the 1970s, when the Council of Europe first raised the possibility of a European credit scheme for language learning, related to fixed points in a framework (van Ek, 1975). The first attempt to create such a fixed point was the Threshold Level (van Ek and Trim, 1990), first published in 1985, which described language learning in terms of the social tasks that learners would be expected to carry out at this level.

Over the years the documentation associated with the CEFR project has acquired the language of competence and constructs, but these do not really impact upon the approach adopted, which remains firmly behavioural (Little, 2006). For example, the CEFR claims to be an 'action-oriented approach' that

> *views users and learners of a language primarily as 'social agents', i.e. members of society who have tasks (not exclusively language-related) to accomplish in a given set of circumstances, in a specific environment and within a particular field of action. While acts of speech occur within language activities, these activities form part of a wider social context, which alone is able to give them their full meaning. We speak of 'tasks' in so far as the actions are performed by one or more individuals strategically using their own specific competences to achieve a given result. The action-based approach therefore also takes into account the cognitive, emotional and volitional resources and the full range of abilities specific to and applied by the individual as a social agent.*
>
> (Council of Europe, 2001: 9)

The scales contained in the document contain no references to competences, only to what learners 'can do' in a variety of communicative situations, using functional language. The scales themselves have no basis in analysis of actual performance, or second language acquisition theory (Fulcher, 2004; 2008b; Hulstijn, 2007); and the primary authors of the scales admit that they are atheoretical (North and Schneider, 1998: 242–243). Rather, the verbal descriptors or 'can do' statements are drawn from a compilation of other rating scales (North, 1993), which were given to teachers to decide which were relevant to their own teaching context in Europe. From their experience, teachers were then asked to place the descriptors in an order from most to least difficult for learners to achieve. Those that could be scaled (that the teachers agreed on most) were used to create the scales in the CEFR. Therefore, as North (2000: 573) says, 'what is being scaled is not necessarily learner proficiency, but teacher/raters' perception of that proficiency – their common framework'. The 'common' in the title therefore refers to the shared perceptions of the teachers who participated in the study to scale the descriptors.

The lack of construct content, the focus on successful behavioural outcomes, and the relative incoherence of the situations across levels of the CEFR, can be seen in most of the scales. In order to illustrate this we reproduce the global scale in Figure 4.6.

Scales of this kind have a superficial attraction. On the surface, the ascending levels appear to increase in difficulty. Indeed, this is how they were created, through perceptions. However, there is little beyond gross observational categories, either in terms of abilities or specific tasks – or 'discursive practices'.

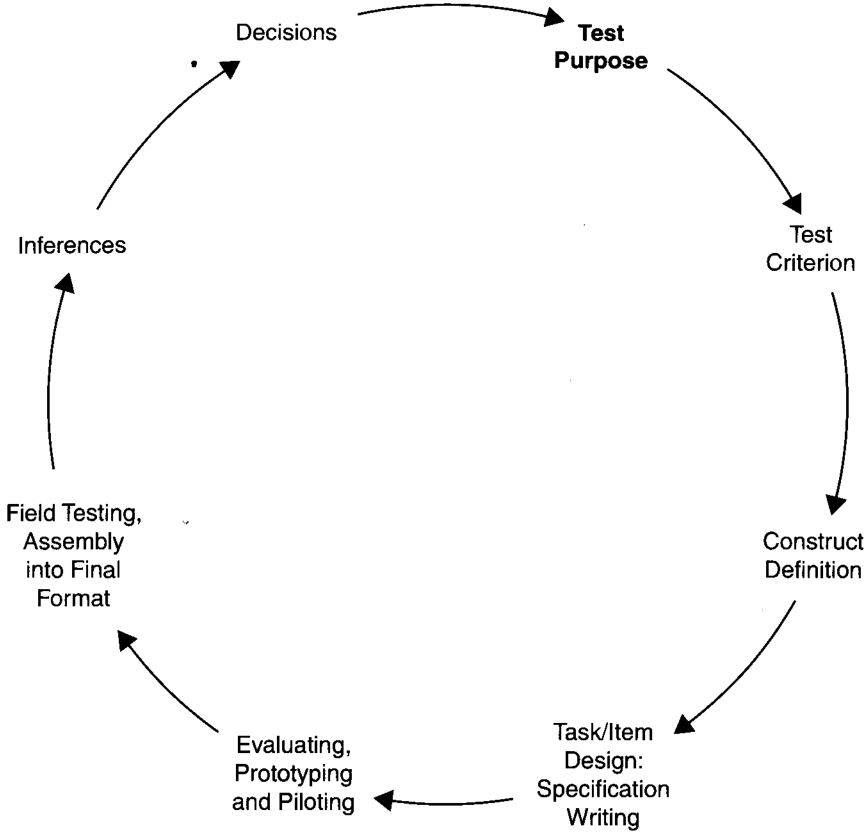
Decisions
Test Purpose
Test Criterion
Construct Definition
Task/Item Design: Specification Writing
Evaluating, Prototyping and Piloting
Field Testing, Assembly into Final Format
Inferences

Fig. 4.6. The common reference levels: global scale

This discussion throws up an interesting question. Why is the Common European Framework called a 'framework', if it is a performance model? We have argued in this chapter that a test framework selects constructs from models and argues for their relevance to a particular purpose for testing. In fact, some people treat the CEFR as if it were really a framework – they take the scales and attempt to use them directly to score actual tests. This is a misuse of the CEFR, which we discuss in Chapter 8. It is indeed an abstract behavioural model, but it acquired its name for bureaucratic reasons. Trim (in Saville, 2005: 282–283) reports that the original intention was to call the CEFR the Common European Model. This name was vetoed by the French representatives in the Council of Europe because for them the word 'modèle' implied an ideal or perfect representation (of language use), and so the term 'framework' was adopted as a compromise. However, in English the term 'model' does not carry these implications; on the contrary, the implication is that it is a poor copy of reality, but may be good enough for its intended purpose. Perhaps in the same way we know that a map of the London underground bears no relationship to the geographical locations of the stations and the tunnels, but that does not stop it from being useful. However, the adoption of the term 'framework' has misled many users into thinking that it is directly applicable to teaching and testing.

This raises the question of just how a performance model like the CEFR can be useful. The answer is that, like all models, it is an attempt, however atheoretical, to describe language use. The descriptors are inadequate, and the levels are flawed because they are not based on primary data. But the text contains ideas that we might take away and use to build a framework for a particular context that is useful. It is a mine, out of which we may chip useful nuggets that, with crafting, can become useful instruments for testing and assessment. One example of how this might work is provided by Davidson and Fulcher (2007). They take the CEFR scale for 'transactions to obtain goods and services', which is reproduced in Appendix 4, and ask what ideas can be used to create a test for successful service encounters. It starts with the two key phrases in the CEFR that seem central to these types of encounters:

- 'Can ask people for things and give people things.'
- 'Can handle numbers, quantities, cost and time.'

As the CEFR does not provide any context or level of complexity, the study draws on discourse analysis of service encounters to create a basic template for a service encounter exchange so that a service encounter 'script' might be assessed at lower ability levels. The assumption for this decision is that one of the first things that learners of a second language need to do is get basic services, like buying a beer or a bus ticket.

Here are two sample items:

Item 1.
[The test taker hears]

Voice 1: Can I buy some apples?
Voice 2: Yes, They're two for 75p.

[The test taker sees]
What comes next?

(a) How much are they?
(b) How much are two?
(c) Thank you. I'll buy two.
(d) Thank you. How much?

Item 2.
[The test taker hears]

Voice 1: By when will my shoes be repaired?
Voice 2: Next Tuesday afternoon, I should think.

[The test taker sees]
What comes next?

(a) Thank you; I'll return Wednesday.
(b) Thank you; I'll return before then.
(c) Will they be ready by Tuesday?
(d) Can I get them on Wednesday?

While using the CEFR in this way can generate test purpose, a context and items, we note that it is still atheoretical. Its use can lead us to treat language use as a purely behavioural phenomenon. This is why further research is needed to supplement ideas mined from the CEFR. For example, in the Davidson and Fulcher study, it was noticed that a key construct in the literature, not mentioned in the CEFR, is the 'ability to establish rapport'. The construct 'rapport' is extremely difficult to define, and the most concise definition comes from the marketing literature:

> *Rapport is a customer's perception of having an enjoyable interaction with a service provider employee, characterized by a personal connection between the two interactants.*
> (Gremler and Gwinner, 2000: 92)

This raises a completely new set of questions relating to the kinds of constructs that might underlie successful service encounters, which may impact not only on testing, but providing both language courses and professional training for staff in a range of service industries. Such testing and training may offset the kinds of problems that have been studied by Ryoo (2005). It seems that one construct of primary relevance is the pragmatics of politeness. This could be operationalised in the talk of either the service provider or the customer, as in this next item, which is a modified version of Item 1 above. In this item, distractor (d) is possible and feasible, in Hymes' terms. It is therefore an alternative to the correct (c), but it is not pragmatically appropriate in this context if rapport is to be maintained.

Item 3.
[The test taker hears]

Voice 1: Can I buy some apples?
Voice 2: Yes. They're two for 75p.

[The test taker sees]
What comes next?

(a) How much are they?
(b) How much are two?
(c) Thank you. I'll buy two.
(d) Gimme two.

Behavioural models must be treated with care. It is not possible simply to use the content for practical language testing purposes. The content and scales need a great deal of applied linguistic work if they are to be applied successfully, and we can see that this work quickly leads us back to a consideration of theoretical constructs.

5. From definition to design

In this chapter we have looked at the test design cycle, and moved along the first three steps from defining test purpose, to construct definition. The amount of time that we spend on the tasks of defining purpose, identifying the criterion and teasing out the constructs that underlie performance will always depend upon the kinds of decisions we wish to make, and how serious those decisions are for the lives of the test takers. In large-scale, high-stakes tests, the process could take months, if not years. For classroom assessments it may be a single meeting of the staff involved. However, I would argue that even for the least important assessments, thinking through the issues in a structured process is professionally rewarding. It brings additional meaning to our understanding of just what it is we are testing, and so to what it is that we think we are teaching. The process of discussion and agreement is one that binds teachers together in a collaborative endeavour to improve teaching and assessment in an environment that encourages professional engagement and development.

While looking at the first three steps in the test development cycle we have also called upon the metaphor of architecture. We have tried to show that there are three levels of architectural design for any test or assessment. In this chapter we have been primarily concerned with the top, or most abstract, level. This is the level of models. We have described these as either construct-based models, or as behavioural-performance models. While we believe that the former are the most productive in test design, we have also tried to show that the latter can be useful if treated as heuristic devices to get us to the point where we have more to go on than the models themselves contain. They can act as starting points for our own ideas and research. But, ultimately, this is what all models are about. We attempt to select from them the constructs or other information that we think may be useful to our own situation.

It is describing our own testing purpose and context, and providing a rationale for the relevance of the constructs that we have selected to use, that constitutes the test framework. However, we have also seen that, in the process of constructing the test framework – of moving between the models and the framework – we may very well generate our own constructs that are particularly relevant to the context. These may be linguistic or, in the case of 'rapport', both linguistic and non-linguistic. We only considered the realisation of politeness in linguistic terms, but in a simulation task for more advanced learners we may wish to consider the role of body language and eye contact. These considerations open up both testing and language teaching to the wider world of communication in the criterion context.

Only at the very end of the chapter did we look at three test items. The test items or tasks, and the format of the test, are the lowest, or most specific level, in the test architecture. In Figure 4.2 these are called the test specifications. We turn to describing the test specifications and how they are constructed in the next chapter.