Chapter One

Approaches to Language Test Design: A Critical Review

1.1 Introduction

To help decide on the most suitable formats for inclusion in a test, it is useful to be aware of the alternative approaches to language testing and their limitations in terms of the criteria of validity, reliability and efficiency.

Validity is concerned with whether a test measures what it is intended to measure. Reliability is concerned with the extent to which we can depend on the test results. Efficiency is concerned with matters of practicality and cost in test design and administration.

These glosses should be sufficient to follow the review of approaches to language testing in this chapter. Readers requiring a more detailed treatment before continuing are referred to Chapter Two where a full discussion of these concepts can be found.

Davies (1978) argued that by the mid-1970s, approaches to testing seemed to fall along a continuum stretching from 'discrete' item tests at one end, to integrative tests such as cloze at the other. He took the view that in testing, as in teaching, there was a tension between the analytical on the one hand and the integrative on the other, and considered that (p. 149) 'the most satisfactory view of language testing and the most useful kinds of language tests, are a combination of these two views, the analytical and the integrative.' He went on to say that it was probable, in any case, that no test could be wholly analytical or integrative (p. 149):

> The two poles of analysis and integration are similar to (and may be closely related to) the concepts of reliability and validity. Test reliability is increased by adding to the stock of discrete items in a test; the smaller the bits and the more of these there are, the higher the potential reliability. Validity, however, is increased by making the test truer to life, in this case more like language in use.

Oller (1979), on the other hand, felt that testing should focus on the integrative end of the continuum. He made a strong case for following the swing of the testing pendulum away from what Spolsky (1976) had described as the 'psychometric–structuralist era',

or the so-called 'discrete point' approach to testing, to what he termed 'the psycholinguistic–sociolinguistic era': the age of the integrative test.

In the description of these approaches below, they are treated as if they were 'distinct' or 'pure' types. It is recognised that, in practice, most tests contain elements of the discrete and the integrative, either in the test format or the assessment procedures adopted, but, while the distinction between the two is neither real nor absolute, it is nevertheless felt that approaches to testing can be usefully described in terms of the particular focus they represent.

1.2 The psychometric–structuralist era

The clear advantages of testing 'discrete' linguistic points are that they yield data which are easily quantifiable, as well as allowing a wide coverage of items. Tests which focus on 'discrete' linguistic items are efficient and have the usual reliability of marking associated with objectively scored tests, but both the 'discrete point' approach and the various formats employed in it suffer from the defects of the construct they seek to measure.

The problem with this approach to the measurement of proficiency is that it depends on proficiency being neatly quantifiable in this fashion. Oller (1979, p. 212) outlined the deficiencies in terms of the construct validity of a hypothetically pure form of this approach:

> Discrete point analysis necessarily breaks the elements of language apart and tries to teach them (or test them) separately with little or no attention to the way those elements interact in a larger context of communication. What makes it ineffective as a basis for teaching or testing languages is that crucial properties of language are lost when its elements are separated. The fact is that in any system where the parts interact to produce properties and qualities that do not exist in the part separately, *the whole is greater than the sum* of its parts ... organisational constraints themselves become crucial properties of the system which simply cannot be found in the parts separately.

Oller is on fairly safe ground here as most people would probably agree that testing a candidate's linguistic competence is a necessary, but not sufficient, component of a test battery. In another context, people taking a driving test are required to demonstrate that they can perform the task. The licensing authority does not depend solely on a pencil and paper test to inform about the extent of examinees' knowledge concerning the principles of driving. Similarly, those who have to make assessments about a piece of music will make them on the piece as a whole, not on selected parts of it. Chaplen (1970a, p. xxvii) criticised isolated skills tests from this point of view, arguing that: 'It seems unlikely that measurements of the component skills most commonly isolated can provide either singly or in aggregate, a satisfactory measurement of the gestalt.' This is a view shared by Savignon (1972) who found that grammatical competence was not by itself a good predictor of communicative skills.

Kelly (1978) argued that if the goal of applied linguistics was seen as the applied analysis of meaning, e.g., the recognition of the context-specific meaning of an utterance as distinct from its system-giving meaning, then applied linguists should be more interested in the development and measurement of ability to take part in specified communicative performance, the production of and comprehension of coherent discourse, rather than in linguistic competence. This echoed Spolsky's (1968) earlier point that perhaps instead of attempting to establish a person's knowledge of a language in terms of a percentage mastery of grammar and lexis, we would be better employed in testing that person's ability to perform in a specified socio-linguistic setting.

Rea (1978, p. 51) has expressed a similar view:

> although we would agree that language is a complex behaviour and that we would generally accept a definition of overall language proficiency as the ability to function in a natural language situation, we still insist on, or let others impose on us, testing measures which assess language as an abstract array of discrete items, to be manipulated only in a mechanistic way. Such tests yield artificial, sterile and irrelevant types of items which have no relationship to the use of language in real life situations.

Morrow (1979) argued that if we are to assess proficiency, i.e., potential success in the use of the language in some general sense, it would be more valuable to test for a knowledge of and an ability to apply the rules and processes by which these discrete elements are synthesised into an infinite number of grammatical sentences and then selected as being appropriate for a particular context, rather than simply to test knowledge of the elements alone. Morrow (1979, p. 145) argues the point that:

> knowledge of the elements of a language in fact counts for nothing unless the user is able to combine them in new and appropriate ways to meet the linguistic demands of the situation in which he wishes to use the language.

1.3 The psycholinguistic–sociolinguistic era

In response to a feeling that 'discrete point' tests were insufficient indicators of language proficiency, the testing pendulum on the whole swung in favour of global tests in the 1970s, into what Spolsky (1976) termed the psycholinguistic–sociolinguistic era, an approach to measurement that was in many ways contrary to the allegedly atomistic assumptions of the 'discrete point' tests (see Davies, 1978).

It was claimed by Oller (1979) that global integrative tests such as cloze and dictation went beyond the measurement of a limited part of language competence achieved by 'discrete point' tests with their bias towards testing the receptive skills; that such tests could measure the ability to integrate disparate language skills in ways which more closely approximated the actual process of language use. Oller's view (1979, p. 37) was that:

> The concept of an *integrative* test was born in contrast with the definition of a discrete point tests. If discrete items take language skill apart, integrative tests put it back together. Whereas discrete items attempt to test knowledge of language one bit at a time, integrative tests attempt to assess a learner's capacity to use many bits all at the same time, and possibly while exercising several presumed components of a grammatical system, and perhaps more than one of the traditionally recognized skills or aspects of skills [author's italics].

Read (1981a, p x) succinctly described the psycholinguistic–sociolinguistic era

> From a psycholinguistic perspective, language came to be seen as less of a well-defined taxonomic structure and more of a dynamic, creative, functional system It was recognised that natural language contains a considerable amount of redundancy, so that it is difficult to show that any single linguistic unit is indispensable for communication The sociolinguistic contribution centres on the concept of communicative competence, which represents a broadening of Chomsky's notion of competence to cover not only knowledge of rules for forming grammatical sentences but also rules for using those sentences appropriately with different contexts Thus the psycholinguistic and sociolinguistic perspectives have enlarged the basis on which the validity of a test is to be judged New criteria have become introduced that cannot be measured by the standard 'objective' methods

Oller maintained that provided linguistic tests such as cloze require 'performance' under real life constraints, e g , time, they are at least a guide to aptitude and potential for communication, even if they do not test communication itself They are also practicable to administer, economical to set and mark, and have respectable reliability figures associated with them

Work by Alderson (1978a), however, raised serious questions about the validity of these integrative measures as testing devices He demonstrated that there is no such thing as '*the* cloze test' and even in using the same passage, results are affected by altering the point where the deletions are started from, or by using a different n/th rate deletion (The cloze procedure is examined in more detail in Section 4 1 3 where consideration is given to the construction and the potential advantages and disadvantages of this test method)

A major cause of concern is the assumption made by Oller (1976, 1979, 1980) that General Language Proficiency (GLP), the grammar of expectancy his integrative tests are aimed at, is a single principal factor underlying all language skills His concept of 'overall proficiency' has inevitably merged into a hypothesis of an underlying unitary competence

This is a view implicit in Oller's concept of the internalised expectancy grammar and, though it is one which is seductive for the purpose of those having to take administrative decisions, as Davies (1981b) points out, it conflicts with substantial evidence in favour of at least two competences, namely reception and production (see Vollmer, 1981) The differences between knowing how to analyse input and knowing how to construct output would seem to outweigh the correspondence between the two processes Pedagogical experience would also suggest that the different performance tasks an individual is faced with result in a variety of different proficiencies being exhibited in the completion of these tasks

Davies (1981b) emphasised that although Oller has claimed that his integrative tests represent total language proficiency better than any other single test or combination of tests, this is not in itself an argument in favour of the unitary competence hypothesis, as measures such as cloze and dictation are so integrative that they contain most or all language abilities anyway High correlations between cloze and other measures may only reflect that they are measuring different skills which are highly correlated among individuals, however, this does not mean that there will be no individuals whose performances in the various skills differ considerably

A group of examinees may have scores in two tests which correlate very highly, in the sense that both tests put the individuals in more or less the same rank order, but since correlational measures take little or no account of mean scores, the group's scores may be centred on very different means in the two tests, indicating quite different levels of performance overall In other words, correlational data do not provide evidence about standards

The empirical evidence that has been marshalled in favour of the 'unitary competence hypothesis' is open to some doubt and there is a growing body of evidence favouring a divisibility hypothesis (see Vollmer, 1979, Bachman and Palmer, 1981a, Hughes, 1981a, Vollmer, 1981, and Porter, 1983)

Principal component analysis is often used to substantiate the 'unitary competence hypothesis', but this method is essentially designed to simplify data, and would be expected to produce one factor from a battery of seemingly different language tests (see Porter, 1983) More crucially, this general language proficiency factor does not necessarily explain all the variance in the results, and the percentage of variance explained differs from study to study (see Vollmer, 1981) Because of the existence of factors other than the principal component, which explain reasonable proportions of the remaining variance, it is often possible by pursuing further factor analysis, for example Varimax rotation of the factor structure, to obtain a number of independent factors each of which makes a sizeable contribution to the total variance

There is also evidence in the literature that the format of a task can unduly affect the performance of some candidates (see Bonikowska, 1986, Murphy, 1978, 1980, and Weir, 1983a) This makes it necessary to include a variety of test formats for assessing each construct rather than rely on a single overall measure, such as cloze

Though the tests Oller has advocated are global in that they require examinees to exhibit simultaneous control over different aspects of the language system, they are nevertheless indirect Although the tests might integrate disparate language skills in ways which more closely approximate actual language use, one would argue that their claim to the mantle of communicative validity remains suspect, as only direct tests which simulate relevant authentic communication tasks can claim to mirror actual communicative interaction (see Kelly, 1978, Morrow, 1979) As Moller (1982b, p 25) pointed out, the indirect tests Oller has advocated do not 'require subjects to perform tasks considered to be relevant in the light of their known future use of the language '

Advocates of communicative language testing would argue that Oller's view pays insufficient regard to the importance of the productive and receptive processing of discourse, arising out of the actual use of language in a social context with all the attendant performance constraints, e g , the interaction-based nature of discourse, unpredictability and behavioural outcomes (see Morrow, 1979, Moller, 1981b) Both Rea (1978) and Morrow (1979) have emphasised that although indirect measures of language abilities claim extremely high standards of reliability and concurrent validity as established by statistical techniques, their claim to other types of validity remains suspect

Morrow (1979) cited as evidence for this the fact that neither cloze nor dictation offers the opportunity for spontaneous production by the candidate and the language norms which are followed are those of the examiner (or original author of the text), not of the student himself Neither testing procedure offers the possibility for oral or non-controlled written

production and since the oral and written skills are generally held to be highly important, some means of assessing them reliably in communicative situations should be found. Although integrative measures appear to correlate highly with other similar measures of general language proficiency, there is empirical evidence that cloze correlates only moderately with tests of written production (see Weir and Ormiston, 1978) and with spoken production (see Vollmer, 1981). Given that the tests concerned are reliable, this would suggest the possibility that proficiency in these areas cannot be adequately predicted by a test of overall proficiency.

Morrow also claimed both cloze and dictation to be fundamentally suspect since they are tests of underlying ability (competence) rather than actual performance. In other words, they depend basically on a knowledge of the language system rather than the ability to operate this sytem in authentic settings. B.J. Carroll (1980b, p. 9) reached the same conclusion: 'this (cloze test) is still essentially usage based. The task does not represent genuine interactive communication and is, therefore, only an indirect index of potential efficiency in coping with day-to-day communicative tasks.'

Even if it were decided that indirect tests such as cloze were valid in some sort of derived fashion, it still remains true that performing on a cloze test is not the same sort of activity as reading. The pedagogical consequences of including this type of test measure in a battery might be harmful if it results in candidates being taught specifically to handle indirect assessment tasks in preference to teaching them to cope with more realistic tasks.

Kelly (1978, p. 241) made the further point that some candidates may manage to succeed in the indirect task by training of a certain kind and thus invalidate the test: 'indirect tests are subject to attacks on their validity in those cases where it is possible to bypass the ability in question and develop proficiency in the assessment task alone.' He also noted (1978, pp. 245–6) that:

> Analysis of a student's responses to an indirect test will not provide any relevant information as to the reasons for the student's difficulties in the authentic task, of which one assumes, the indirect test is a valid and reliable measure. By their very nature, indirect tests can provide evidence for *level* of achievement, but cannot diagnose specific areas of difficulty in relation to the authentic task.

Integrative tests such as cloze only tell us about a candidate's linguistic competence. They do not tell us anything directly about a student's performance ability, and their main value in their unmodified form is in designating competence levels rather than relating candidates' performance to any external criteria. They are perhaps only of limited use where the interest is in what the individual student can or cannot do in terms of the various language tasks he may face in real life situations.

The deficiencies in the type of information the 'discrete point' approaches of the psychometric–structuralist era and the more integrative approaches of the psycholinguistic–sociolinguistic era can provide bring about a need to investigate the 'communicative paradigm' to see whether this approach might prove more satisfactory.

1.4 The communicative paradigm

1.4.1 Terminology

There is a potential problem with terminology in some of the literature on communicative approaches to language testing. References are often made in the literature to testing communicative 'performance', e.g., B.J. Carroll's book (1980b) is entitled *Testing communicative performance*. It seems reasonable to talk of testing performance if the reference is to an individual's performance in one isolated situation, but as soon as we wish to generalise about ability to handle other situations, 'competence' as well as 'performance' would seem to be involved, or more precisely 'capacity' in the Widdowsonian sense (Widdowson, 1983). Bachman's use of the term 'communicative language ability' which includes both knowledge, or competence, and the capability for implementing that competence in language use would seem to be consistent with Widdowson's term in providing a more inclusive and satisfactory definition of language proficiency.

Strictly speaking, a performance test is one which samples behaviour in a single setting with no intention of generalising beyond that setting — otherwise a communicative language test is bound to concern itself with 'capacity' (Widdowson, 1983) or 'communicative language ability' (Bachman, 1990). The very act of generalising beyond the setting actually tested, implies some statements about abilities to use the language and/or knowledge of it. Conversely it is difficult to see how competence (knowing about using a language) might be evaluated except through its realisation in performance. Only performance can be directly observed and hence evaluated. All linguistic behaviour, even completing multiple choice tests of phoneme discrimination, necessarily involves performance. In practice a clear distinction between performance and competence will be difficult to maintain.

In testing communicative language ability we are evaluating samples of performance, in certain specific contexts of use, created under particular test constraints, for what they can tell us about a candidate's communicative capacity or language ability. Skehan (1988) points out that while such tests may not replicate exactly the performance conditions of a specific task in the target situation they are likely to replicate to some degree conditions of actual performance.

Skehan summarises the current position succinctly:

> What we need is a theory which guides and predicts how an underlying communicative competence is manifested in actual performance; how situations are related to one another, how competence can be assessed by examples of performance on actual tests; what components communicative competence actually has; and how these interrelate .... Since such definitive theories do not exist, testers have to do the best they can with such theories as are available.

1.4.2 The theoretical base

The validity of tests which claim to be communicative is a function of the degree of understanding of communication and communicative ability on the part of the test constructor. It is naïve to asssume that one can develop valid tests of communicative language ability without reference to the construct which one is attempting to measure,

arguments relating to the state of the available descriptions of language in use not withstanding

Agreement on what components should be included in a model of communicative language ability is by no means unanimous (see Courchene and de Bagheera, 1985, p 49) Indeed relatively little is known about the wider communicative paradigm in comparison with linguistic competence per se and adequately developed theories of communicative language use are not yet available This is not to say we must wait for completion of such theories before appropriate testing procedures can be developed Rather we need to investigate systematically some of the available hypotheses about language use and try to operationalise these for testing purposes In this way the constructs and processes of applied linguistics may be examined empirically and their status evaluated

Canale and Swain (1980) provided a useful starting point for a clarification of the terminology necessary for forming a more definite picture of the ability to use language communicatively These authors took communicative competence to include grammatical competence (knowledge of the rules of grammar), sociolinguistic competence (knowledge of the rules of use and rules of discourse) and strategic competence (knowledge of verbal and non-verbal communication strategies) The model was subsequently updated by Canale (1983), who proposed a four-dimensional model comprising linguistic, sociolinguistic, discoursal and strategic competences, the additional distinction being made between sociolinguistic (sociocultural rules) competence and discoursal competence (cohesion and coherence)

The framework proposed by Bachman (1990) is consistent with these earlier definitions of communicative language ability

> Communicative language ability consists of language competence, strategic competence, and psychophysiological mechanisms Language competence includes organisational competence, which consists of grammatical and textual competence and pragmatic competence, which consists of illocutionary and sociolinguistic competence Strategic competence is seen as performing assessment, planning and execution functions in determining the most effective means of achieving a communicative goal Psychophysiological mechanisms involved in language use characterise the channel (auditory, visual) and mode (receptive, productive) in which competence is implemented

Language competence is composed of the specific knowledge and skills required for operating the language system, for establishing the meanings of utterances, for employing language appropriate to the context and for operating through language beyond the level of the sentence Strategic competence consists of the more general knowledge and skills involved in assessing, planning and executing communicative acts efficiently Skehan (1988) suggests that the strategic component is implicated when communication requires improvisation because the other competences are in some way insufficient The final part of Bachman's model deals with skill and method factors which are meant to handle the actual operation of language in real situations and so locate competence in a wider performance framework

Models such as these provide a potentially useful framework for the design of language tests, but it must be emphasised that they are still themselves in need of validation (see

Brindley, 1986, Swain, 1985) The existence of the components of the model even as separate entities has not been established Skehan (1988) rightly points out that the relationship between the various competences is not entirely clear, nor is the way they are integrated into overall communicative competence Nor is it made clear how this communicative competence is translated into communicative performance Candlin (1986) also outlined some of the problems to be faced in testing communicative competence and argued that their solution depends first on our description of this construct

To date a limited amount of research has been carried out on investigating the measurement of language competence and method factors but very little has been done on the specific measurement of communication strategies or its relationship to the other competences This in itself may be an indication of the inherent difficulties in this area There is a pressing need for systematic research to illuminate all of these unresolved issues

To help clarify what is meant by communicative testing we are forced to resort to available pretheoretical data from the literature relating to the concept of communicative competence Since Hymes's two-dimensional model of communicative competence, comprising a 'linguistic' and a 'sociolinguistic' element, most subsequent models have included consideration of a sociolinguistic dimension which recognises the importance of context to the appropriate use of language and the dynamic interaction that occurs between that context and the discourse itself

For Hymes (1972), communicative competence had included the ability to use the language, as well as having the knowledge which underlay actual performance Morrow (1979) felt that a distinction needed to be made between communicative competence and communicative performance, the distinguishing feature of the latter being the fact that performance is the realisation of Canale and Swain's (1980) three competences and their interaction 'in the actual production and comprehension of utterances (under general psychological constraints that are unique to performance) ' (Morrow, 1979)

Morrow (1979) and Canale and Swain (1980) argued that communicative language testing, as well as being concerned with what the learner knows about the form of the language and about how to use it appropriately in contexts of use (competence), must also deal with the extent to which the learner is actually able to demonstrate this knowledge in a meaningful communicative situation (performance), i e , what he can do with the language, or as Rea (1978, p 4) put it 'his ability to communicate with ease and effect in specified sociolinguistic settings '

The capacity or ability (see Widdowson, 1983, Bachman, 1990) to use language communicatively thus involves both competence and demonstration of the ability to use this competence It is held that the performance tasks candidates are faced with in communicative tests should be representative of the type of task they might encounter in their own real-life situation and should correspond to normal language use where an integration of communicative skills is required with little time to reflect on, or monitor language input and output The criteria employed in the assessment of performance on these tasks should relate closely to the effective communication of ideas in that context

This perspective is consistent with the work of language testers generally supportive of a broadly based model of communicative language ability where there is a marked shift in emphasis from the linguistic to the communicative dimension The emphasis is no longer

on linguistic accuracy, but on the ability to function effectively through language in particular contexts of situation.

Cooper's (1968) view that existing test frameworks, because they concentrated on linguistic competence, might fail to assess a person's communicative ability, was taken up by Morrow (1979, p. 149) who argued that traditional tests did not give:

> any convincing proof of the candidate's ability to actually use the language, to translate the competence (or lack of it) which he is demonstrating into actual performance 'in ordinary situations', i.e., actually using the language to read, write, speak or listen in ways and contexts which correspond to real life.

B.J. Carrol (1989b, p. 1) adopted a similar line:

> the prime need of most learners is not for a theoretical or analytical knowledge of the target language, but for an ability to understand and be understood in that language within the context and constraints of particular language-using circumstances.

His opinion (1989b, p. 7) is that: 'the ultimate criterion of language mastery is therefore the learner's effectiveness in communication for the settings he finds himself in.'

Rea (1985) argued that all tests can be seen as tests of performance that are to varying degrees communicative or non-communicative, or (in Widdowson's dichotomy) use- or usage-based. Rea further distinguishes between items as meaning-dependent or meaning-independent, and describes how the former can be subdivided according to whether they involve a context determined response or not.

Moller (1981b) felt that communicative performance relates to the transmission and reception of particular meanings in particular contexts, and what can be tested is the quality and effectiveness of the performance observed in these circumstances. Kelly (1978, p. 350) had argued in a similar vein:

> To take part in a communicative event is to produce and/or comprehend discourse in the context of situation and under the performance conditions that obtain. It is the purpose of a proficiency test to assess whether or not candidates are indeed capable of participating in typical communication events from the specified communication situation(s).

These statements reflect an emphasis in language teaching and, more recently, testing that has been placed on *use* and the concern that has been shown with communicative functions rather than with the formal language patterns of *usage* (see Campbell and Wales, 1970; Hymes, 1972 and Widdowson, 1978 and 1983). Such theoretical descriptions are essential for describing the broad parameters within which communicative language testing should fall but practitioners need more tangible attributes to ascertain the degree of communicativeness of a test or to make their tests as communicative as possible within the constraints obtaining. What does a communicative test look like? How does it differ from other tests? These are questions which we need to address.

1.4.3 Distinguishing features of communicative language tests

Only a few of the currently available theories of language use seem amenable to the demands of language testing. It is therefore essential to be as precise as possible about the skills

and performance conditions for any tests which claim to assess communicative language ability. Test constructors must closely identify those skills and performance conditions (see Skehan, 1988) that are the most important components of language use in particular contexts. The incorporation of these features, where appropriate, would indicate the degree to which the test task reflected the attributes of the activity in real life that it was meant to replicate. Unless steps are taken to identify and incorporate such features it would seem imprudent to make statements about a candidate's ability to function in normal conditions in his or her future target situation.

We also have to ensure that the sample of communicative language ability in our tests is as representative as possible. What and how to sample with our tests is a key issue in language testing. Many of the available descriptions of use are now both detailed and quite extensive, but are not always called on by testers. We need to make good this deficiency. If we are to extrapolate from our test data and make statements about communicative language ability in real life situations, great care needs to be taken over the texts and tasks we employ in our tests. These should accord as far as possible with the general descriptive parameters of the intended target situation particularly with regard to the skills necessary for successful participation in that situation. Additionally, tests should meet the performance conditions of that context as fully as possible. The difficulties of achieving this match with real life and the resultant implications for generalisability from communicative test data are discussed in the final section of this chapter.

In the testing literature there is a strong emphasis on the importance of test purpose, and it is held that no one solution can accommodate the wide variety of possible test scenarios. It is argued that appropriately differentiated tests in different skills areas need to be made available for evaluating different groups of examinees with different target situation needs. To measure language proficiency adequately in each situation, account must now be taken of: where, when, how, with whom, and why the language is to be used, and on what topics, and with what effect. The fact that communicative performance is affected by prior knowledge/experience/abilities is accepted along with the implications of this for test specificity (see Alderson and Urquhart, 1985b).

The important role of context as a determinant of communicative language ability is stressed and an integrative approach to assessment as against a decontextualised approach is advocated. Language can not be meaningful if it is devoid of context (linguistic, discoursal and sociocultural). For Oller (1973, 1979) the higher the level at which language is contextualised, the more effective language perception, processing and acquisition are likely to be. The variability in performance, according to the discourse domain or type of task involved, is recognised with the attendant implications this might have for test length and the types of text and formats to be included in a test battery (see Douglas and Selinker, 1985; Skehan, 1987).

The authenticity of tasks and the genuineness of texts in tests is regarded as something worth attempting to pursue despite the problems involved both in the definition of this and in its realisation. If inauthentic tasks are included in tests of communicative language ability there is a real danger that the method employed could interfere with the measurement of the construct we are interested in. We could end up measuring ability to cope with the method rather than the ability to read, listen, write, speak or deal with a combination

of these skills in specified contexts. The more authentic the tasks the less we need to be concerned about this. If certain techniques only occur in tests, e.g., cloze or multiple choice, why should we ever contemplate their use? Tests of communicative language ability should be as direct as possible (attempt to reflect the 'real life' situation) and the tasks candidates have to perform should involve realistic discourse processing.

Unsimplified language, i.e., non doctored, 'genuine' texts should be used as inputs (see Widdowson, 1983) and due reference made to the referential and functional adequacy of these. In addition attention needs to be paid to other task dimensions such as the size of the text to be understood or produced and to processing in real time.

The net result of these considerations is that different tests need to be constructed to match different purposes and these instruments are no longer uniform in content or method. A variety of tests is now required whereas within previous orthodoxies authors were satisfied with a single 'best test'.

In assessing the ability to interact orally we should try to reflect the interactive nature of normal spoken discourse and attempt to ensure that reciprocity is allowed for in the test tasks included. The tasks should be conducted under normal time constraints and the element of unpredictability in oral interaction must be recognized, for authentic communication may lead the participants in unforeseen directions. Candidates may also be expected in certain tasks such as group discussion to demonstrate an ability to manage the interaction and/or to negotiate meaning with interlocutors. In short what we know from the theory of spoken interaction should be built into tasks which purport to test it (see Bygate, 1987 and Weir and Bygate, 1990).

The legitimacy of separate skills testing is being questioned, however, and indeed the more innovatory testing of skills through an integrated story-line set of procedures (see Low, 1986) is gaining favour. The discredited holy grails of the psycholinguistic–sociolinguistic era, such as cloze, are still seen to have a minor role to play in adding to the reliability of test batteries and assessing the more specifically linguistic skills, but centre stage is now given to more direct attempts to operationalise the integrated testing of communicative language ability.

Direct testing requires an integrated performance from the candidate involving communication under realistic linguistic, situational, cultural and affective constraints. Candidates have to perform both receptively and productively in relevant contexts. The focus is on the expression and understanding of functional meaning as against a more limited mastery of form. The move to direct testing has been further encouraged by a concern among language testers about the problems of format effect.

Format effect relates to the possibility that test results may be contaminated by the test format employed, i.e., a different estimate of a skill such as reading might be obtained if a different format is employed. This possible influence of test method on trait estimation is increasingly recognised if not yet fully understood (see Bachman and Palmer, 1982; Bachman, 1990). There is some evidence (Murphy, 1978, 1980; Weir, 1983a) to suggest that multiple choice format may be particularly suspect in this respect.

In order to elicit the student's best performance it is important to minimise any detrimental effect of the techniques of measurement on this performance. It is felt that the type of performance elicited by certain assessment methods may be qualitatively different from

real life language use and to the extent that this is the case it is difficult to make statements about candidates' language proficiencies.

In the area of marking, the holistic and qualitative assessment of productive skills, and the implications of this for test reliability, need to be taken on board. The demands of a criterion-referenced approach to testing communicative language ability (where statistical analysis is likely to be more problematic) and the establishment of meaningful cut off scores demand attention (see Brindley, 1986; Cziko, 1981; and Hauptman *et al.*, 1985). At the final stage of the testing process the profiling of test results has to be addressed as we abandon the notion of a single general proficiency.

Though it is accepted that linguistic competence must be an essential part of communicative competence, the way in which they relate to each other, or indeed how either relates to communicative ability (in the performance sense), has not been clearly established by empirical research. A good deal of work needs to be done in comparing results obtained from performance on non-communicative, linguistically-based tests with those which sample communicative ability through instances of performance, before one can make any positive statements about the former being a sufficient indication of likely ability in the latter or in real life situations. No realistic comparisons are possible until reliable and effective, as well as valid, methods are investigated to assess proficiency in performing relevant communicative tasks.

For testers operating within the communicative paradigm there is greater pressure to validate tests because of an expressed desire to make the tests as direct as possible both in terms of task and criteria. Claims made for tests being able to measure or predict real life language performance adequately must be tentative until the validity of the measures used is substantiated. There is a pressing need to establish the theoretical and empirical validity of measures conceived within this paradigm.

The use of introspection studies to investigate the validity of a skills approach to the testing of reading is an example of one way validation studies might develop in the future. Candidates taking a reading test might be asked to verbalise how they answer each item and the results of these investigations could then be compared with the tester's intentions in setting each item. This might shed light on the nature of the reading construct itself and the way suspected component skills relate to each other (or not) (see also Candlin, 1986).

The commitment to making tests communicative thus entails a high degree of explicitness both at the test design stage where one is concerned with the required result and at the evaluation stage where one is estimating the acquired result (see Hawkey, 1982). It is not necessarily the case that communicative tests will look radically different from some existing tests; but there may be strong pragmatic reasons for trying to demonstrate any difference in either the test content, the marking schemes to be applied and the way results are reported.

In the present state of uncertainty the effect of the test on the teaching that precedes it should receive serious consideration. If our communicative tests have a beneficial backwash effect in encouraging the development of communicative capacity in the classroom (see Swain, 1985; Hughes 1989) then we can be less worried about the theoretical or empirical shortcomings of our knowledge of language in use. Similarly if we can include

in our tests what is considered to be most appropriate and best practice from the language classroom the match between teaching, testing and reality is that much enhanced. The procedures adopted by the Royal Society of Arts/UCLES in the design of their Certificates in Communicative Skills in English are worthy of note in this respect (see Appendix III).

1.4.4 An acceptable paradigm?

In a recent Test of English as a Foreign Language (TOEFL) conference entitled 'Toward Communicative Competence Testing', the theme was to explore ways in which the TOEFL test might be made more 'communicative' without seriously impairing its present psychometric attributes (see Stansfield, 1986). Bachman (1986) investigated the lack of context associated with many of the TOEFL items and concluded (p. 86) that: 'the majority of the tasks measure only grammatical competence ... with only a handful tapping illocutionary or sociolinguistic competence.'

Douglas (1986) argued for learning from interlanguage studies which had shed light on the variability of performance occasioned by the elicitation procedures employed (task) and by the discourse domain (context) in which the tasks are carried out (see Douglas and Selinker, 1985; Selinker and Douglas, 1985; and Skehan, 1984, 1987). Douglas argued that if the TOEFL and the Test of Spoken English (TSE) were revised in the direction of domain-specific tasks they would fit more easily into a framework of communicative competence. Consideration was also given to the authenticity of the language used (see Bachman 1986; Douglas, 1986) and it was agreed that in the then current TOEFL not enough attention was given (in the listening section) to replicating the normal features of spontaneous spoken discourse, e.g., hesitations, nor to incorporating normal features of interaction such as the negotiation of shared meaning.

What is significant about these discussions at the Educational Testing Service (ETS) at Princeton is that an organisation which had hitherto been operating firmly in the psychometric–structuralist tradition was now concerned in making its tests more communicative. The conference was a limited indication of acceptance of the principles of communicative language testing.

In the language classroom, however, the majority of commercially available tests are still predominantly structure-based (see Archer and Nolan-Woods, 1976; Fowler and Coe, 1978; and Allen, 1982). Most language teaching coursebooks and accompanying teacher manuals, if they contain any advice on testing at all, usually offer vague theoretical generalisations far removed from the practical needs of the teacher who has to construct achievement tests for use in the classroom. Equally unpalatable is the outdated and overly specific advice that is sometimes provided on various discrete point, non-communicative, atomistic approaches, which pay scant regard to any of the insights gained through testing research in the last two decades.

Very little help is normally provided in relating test task to test purpose or in selecting appropriate formats for testing more communicatively. There is an almost total silence on how to interpret test results once the data has been generated.

There is an urgent need for ELT publishers to take account of the developments in the field of communicative language teaching and testing if new initiatives are not to be stifled.

The promising field of Action Research has much to offer in this respect. Cumulative, informal, small-scale investigations by teachers, in a variety of classroom contexts, could help advance our understanding of a whole range of communicative techniques for assessing language proficiency (see Brindley, 1989).

There are strong arguments for following the lead of the UCLES/RSA in this respect. In the design of their certificates in Communicative Skills in English, the test constructors drew heavily on what EFL teachers thought to be sound practice in the classroom. The communicative tests that resulted are in essence classroom-proven teaching techniques which are convertible to elicitation techniques in a test situation (see Appendix III for a full description of this test).

The only necessary difference between teaching and testing within the communicative paradigm relates to the amount of help that is available to the candidate from his teacher or his peers. The help that is normally available in the teaching situation e.g. prompts, reformulation of questions, encouragement, correction and the opportunity to try again, is removed in a test for reasons of reliability of measurement. In this sense the test might be viewed as an intermediate stage between the world of the classroom and the future target situation where the candidate will have to operate unaided.

1.4.5 The promised land?

So far in this chapter the development of a communicative approach to testing language has been outlined and, in general, a positive view has been taken of these developments. There are, however, a number of outstanding problems in adopting an approach within this paradigm that need to be addressed. In order to try and draw these problems together they are considered in relation to the central issue of generalisability of test results. This is an unavoidable issue, whatever approach to testing we adopt.

1.5 The problem of extrapolation

Other than serious marker reliability problems, associated with the assessment of performance (see Section 4.3), the major issue affecting an adoption of a 'communicative' approach to language testing is the generalisability of the results produced by a test.

Any test can be seen as a sampling instrument that provides evidence on which to base inferences that extend beyond the available data. For most purposes the evidence provided by test performances has to be relevant to the whole domain of interest, that is, the test has to be valid; it also has to be capable of allowing stable predictions to be made about a candidate's performance in any part of the domain, in other words, the test has to be reliable.

A communicative test implies the specification of performance tasks closely related to the learner's practical activities, that is, to the communicative contexts in which he would find himself. This creates a problem of generalisability of tasks to be selected. For Kelly (1978, p. 225) the possibility of devising a construct valid proficiency test, i.e., one that measured ability to communicate in the target language, was dependent on the prior

existence of 'appropriate objectives for the test to measure'

There is an often expressed demand in the literature on performance-based tests for a systematic and thorough specification of the communicative demands of the target situation (see Wesche, 1985) Advocates of performance-based tests (see Morrow, 1977, 1979, Carroll, 1980b, 1981b, and Wesche, 1981) seem to be arguing that it is only necessary to select certain representative communication tasks, as we do not use the same language for all possible communication purposes In the case of proficiency tests, these tasks are seen as inherent to the nature of the communication situation for which candidates are being assessed Caution, however, demands that we wait until empirical evidence is available before making such confident statements concerning the identification of these tasks It is only after examining if it is feasible to establish suitable objectives, through empirical research based on real people coping with real situations, that there would be any grounds for claiming to have selected a representative sample of operational tasks to assess performance ability Even when empirical research is conducted to establish suitable objectives, viz , to identify relevant communicative tasks and underlying constituent enabling skills for a target population, the problems of sampling, practicality, reliability and validity still remain

The problems associated with establishing such specifications empirically, the methods to use and the ranking of needs are discussed by Wesche (1981) and Weir (1983a) A factor that emerges clearly is that the very increase in specificity brought about by the needs analysis (particularly of the Munby variety) in itself serves to decrease the possibility of generalisability The more specific the tasks one identifies the less one can generalise from performance on its realisation in a test This type of needs analysis is in any case unable to specify the relative importance of the variables If, as Rea (1978) and Morrow (1979) suggest, the aim should be to construct simulated communication tasks which closely resemble those a candidate would face in real life and which make realistic demands on him in terms of language performance behaviour, it might be difficult to do so reliably or validly Communication is not conterminous with language and much communication in non-linguistic Often the conditions for actual real-life communication are not replicable in a test situation, which is unavoidably artificial and idealised and, to use Davies's (1978) phrase, Morrow and others are perhaps fruitlessly pursuing the chimera of authenticity

Further, even if the sample of communicative tasks possessed content and face validity, might it not still lack generalisability in terms of the other communicative tasks which are not included? Are assessments of performance on these tasks made under particular linguistic and social constraints and thus not relatable to competence as 'characteristic abilities'? In other words, if a selection is made, if a sample is taken from a domain, how can it be ascertained that it is an adequate sample?

Kelly (1978, p 226) observed that any kind of test is an exercise in sampling and from this sample an attempt is made to infer students' capabilities in relation to their performance in general

> That is, of all that a student is expected to know and/or do as a result of his course of study (in an achievement test) or that the position requires (in the case of a proficiency test), a test measures students only on a selected sample The reliability of a test in this conception is the extent to which the score on the test is a stable indication of candidates' ability in relation to the wider universe of knowledge, performances, etc , that are of interest

He pointed out (p 230) that even if we had a clear set of communication tasks

> the number of different communication problems a candidate will have to solve in the real world conditions is as great as the permutations and combinations produced by the values of the variables in the sorts of messages, contexts of situation and performance conditions that may be encountered

Thus, on the basis of performance on a particular item, one ought to be circumspect, to say the least, in drawing conclusions about a candidate's ability to handle similar communication tasks

Morrow (1977, p 53) was also aware of the problems of extrapolation He succinctly set out the problem like this

> The very essence of a communicative approach is to establish particular situations with particular features of context, etc , in order to test the candidate's ability to use language appropriate in terms of a particular specification While it is hoped that the procedures discussed will indeed be revealing in those terms, they cannot strictly speaking reveal anything of the candidate's ability to produce language which is appropriate to a situation different in even one respect from that established

Alderson (Alderson and Hughes, 1981, p 59) also accepted that to follow the communicative paradigm one needed to define what it was that students had to do with language in a specific situation or series of situations, but recognized that by specifying performance in this manner 'one might end up describing an impossible variety of situations, which one cannot encompass for testing purposes '

In order to make stable predictions of student performance in relation to the indefinitely large universe of tasks, it would seem necessary to sample candidates' performances on as large a number of tasks as is possible, which conflicts immediately with the demands of test efficiency The larger the sample of tasks and the more realistic the test items, the longer the communicative test will have to be However, as Alderson noted

> it may be that the issue of extrapolation is not (yet) of crucial importance even if we cannot generalise from performance in one situation to performance in a variety of situations, if we can say something about performance in *one* situation, then we have made progress, and if we can say something important about performance in the target situation so much the better Ultimately the student will have to perform, despite the statistical evidence of the relationship between predictor and predicted, or the theoretised relationship between competence and performance

It may not, however, be all that easy to identify this 'one situation' on which to base our predictions Let us take as an example the development of an EAP reading test Here an additional problem for sampling in a test may well occur in the selection of texts on which a candidate has to demonstrate his or her comprehension skills There is some evidence in the literature that a number of students might be disadvantaged by being tested on their comprehension of texts outside their academic field (see Weir, 1983a, Alderson and Urquhart, 1985a, 1985b) The implication of this might be that different tests are

necessary for audiences which are clearly identifiable as different. There is an urgent need for further investigation into language testing for specified purposes.

Morrow (1977) observed that in the case of conventional language tests aimed at measuring mastery of the language code, extrapolation would seem to pose few problems. The grammatical and phonological systems of a language are finite and manageable and the lexical resources can be delimited. The infinite number of sentences in a language is made up of a finite number of elements, and tests of the mastery of these elements are extremely powerful from a predictive point of view. Thus, Davies (1978, p. 225) remarked: 'What remains a convincing argument in favour of linguistic competence tests (both discrete point and integrative) is that grammar is at the core of language learning . . . . Grammar is far more powerful in terms of generalisability than any other language feature.'

Kelly (1978) provides an interesting argument against this viewpoint. It is not known, for example, how crucial a complete mastery of English verb morphology is to the overall objective of being able to communicate in English, or how serious a disability it is not to know the second conditional. We still do not possess what Kelly (1978, p. 17) described as a: 'reliable knowledge of the relative functional importance of the various structures in a language.'

Given this failing, it would seem ill-advised to make any claims about what students should be able to do in a language on the basis of scores on discrete point tests of syntax or lexis. The construct 'ability to communicate in language' involves more than a mere manipulation of certain syntactic patterns with a certain lexical content. In consequence, it would appear that there is still a need to attempt to devise measuring instruments which can assess performance ability.

As a way out of the extrapolation quandary, Kelly (1978, p. 239) suggested a two-stage approach to the task of devising a test that represents a possible compromise between the conflicting demands of the criteria of validity, reliability and efficiency:

> The first stage involves the development of a direct test that is maximally valid and reliable, and hence inefficient. The second stage calls for the development of efficient, hence indirect, tests of high validity. The validity of the indirect tests is to be determined by reference to the first battery of direct tasks. Clearly, where valid and reliable but inefficient tests already exist for the construct in question, then the research strategy calls for the development of efficient, indirect tests whose results correlate highly with those of the existing test.

Thus, retreat from direct evaluation of performance may be acceptable, *provided* relationships or even correlations between data from competence testing and predicted behaviour have been established.

As far as large scale proficiency testing is concerned, another viable solution might be to focus attention on language use in individual and specified situations while retaining, for purposes of extrapolation, tests of the candidate's ability to handle those aspects of language which are generalisable to all language use situations, namely the grammatical and phonological systems. Or it may be the case that the former make the latter unnecessary (see Weir, 1983a) as was the case in the TEEP research, *where the grammar test was abandoned because it offered no additional information to that provided by the more use based components.*

Morrow (1979, p. 152) saw a third way out of the extrapolation quandary. His argument is that a model (as yet unrealised) for the performance of global communicative tasks may show, for any task, the enabling skills which have to be mobilised to complete it:

> The status of these enabling skills vis-à-vis competence : performance is interesting. They may be identified by an analysis of performance in operational terms, and thus they are clearly, ultimately performance-based. But at the same time, their application extends far beyond any one particular instance of performance, and in this creativity they reflect an aspect of what is generally understood by competence. In this way they offer a possible approach to the problem of extrapolation.

He asserted that (p. 153): 'Analysis of the global tasks, in terms of which the candidate is to be assessed, will usually yield a fairly consistent set of enabling skills' and argues that assessment of ability in using these skills would therefore yield data which are relevant across a broad spectrum of global tasks, and are not limited to a single instance of performance.

For Morrow (1979, p. 153), a working solution to the problem would be the development of tests which measure both overall performance in relation to a specified task and the strategies and skills which have been used in achieving it:

> Written and spoken production can be assessed in terms of both these criteria. In task-based tests of listening and reading comprehension, however, it may be rather more difficult to see just how the global task has been completed. . . . it is rather difficult to assess why a particular answer has been given and to deduce the skills and strategies employed. In such cases, questions focusing on specific enabling skills do seem to be called for in order to provide the basis for convincing extrapolation.

He is aware, though, that there exists in tests of enabling skills a fundamental weakness in the relationship between the whole and the parts, as a candidate may prove quite capable of handling individual enabling skills, yet still not be able to communicate effectively.

Another problem is that it is by no means easy to identify these enabling skills; nor are there any guidelines for assessing their relative importance for the successful completion of a particular communicative task, let alone their relative weighting across a spectrum of tasks. Morrow would appear to assume that we are not only able to establish these enabling skills, but also to describe the relationship that exists between the part and the whole in a fairly accurate manner (in this case, how 'separate' enabling skills contribute to the communicative task). He appears to be saying that there is a prescribed formula; possession and ability to use enabling skills X + Y + Z = successful completion of communicative task (1), whereas it would seem likely that the added presence of a further skill or the absence of a named skill might still result in successful completion of the task in hand.

A pragmatic way out of this dilemma of how to know what we are testing would be to pursue an ethnographic validation approach as outlined in Section 2.1.2. Data could be collected from student introspections on the processes they are utilising to complete items. This could be used to help determine which items best fitted the required specification (see Aslanian, 1985; Cohen, 1985; and Jones and Friedl, 1986).

In addition, advice could be taken from professionals who control the content and the language of purposive interactions in the target domain of proficiency test candidates. They could be asked to comment on the appropriateness of test items for the intended populations (see Douglas and Pettinari, 1983). The recent IELTS revision project has adopted this useful strategy.

The extrapolation problem faced by those adopting a more communicative approach to language test design seems to relate to the wider issue of the status of laws in the behavioural sciences. In the physical sciences, laws are extrapolations of replicable phenomena. Researchers in these domains can directly confront what they wish to investigate, formulate hypotheses and repeat experiments as many times as they wish to verify or falsify their hypotheses. Because of problems associated with the infinite variability of language in use and the problems involved in population sampling, the scientific paradigm is a difficult one to follow in educational measurement.

Hawkey (1982) described the classical scientific paradigm as a hypothetico-deductive methodology formulating quantifiable, narrow, parsimonious hypotheses, tested through the observation of the behaviour of a random sample of the target population, followed by a statistical analysis of the results according to pre-ordained procedures. This approach is not suitable for large scale proficiency testing where candidates might be operating in a variety of contexts and at a variety of levels. Account may need to be taken of a large number of variables, some of which are not predictable, all interacting in socio-cultural contexts. Thus there is a task sampling problem, a validity problem.

Unlike the scientific paradigm described by Hawkey there might also be serious problems in terms of population sampling. If the target population of students is transient, widely dispersed and varied in terms of accessibility, the sampling might of necessity have to be opportunistic. This is a population sampling problem, a reliability problem.

The concern might have to be, of necessity, with what Hawkey (1982, p. 16) described as an 'illuminative evaluation' paradigm, where the focus was on the description of complex phenomena, the resolution of significant features, and the comprehension of relationships. Initial research in this area might have to limit itself to providing a descriptive framework for establishing communication tasks of relevance to students in a specified context, prior to test construction (see Weir, 1983b). No definitive claims could be made (without empirical validation studies) about how the language user operates when involved in these communication tasks or how he learns to perform such tasks (see Kelly, 1978). What might be provided is a specification, coarse but robust, of the general communicative tasks facing target students in their specified context.

Irrespective of the problematic nature of the exercise, the need for specifying as clearly as possible what it is that is to be tested seems axiomatic for testing within the communicative paradigm. The current interest in ESP is a reflection of this and the acronym might better be regarded as English for Specified Purposes rather than specific or special. This would emphasise the belief that all teaching and testing is to varying extents specified and never totally general.

However, the nature and shortcomings of target-situation analysis for arriving at specifications of language needs for tests have been discussed extensively in the literature (see Weir, 1983a). There are dangers in analyses being too specific, e.g., they may not

be operationalisable in a test, as well as in being too general, e.g., they may disadvantage certain candidates. Perhaps the biggest danger is that there is a tendency for needs analysis to claim a disproportionate amount of the time and resources available for research, often at the expense of test development.

A comprehensive description of the specification stage of a test design can be found in Weir (1983a, 1983b), where the extensive research and development behind the Associated Examining Board's Test in English for Educational Purposes (TEEP) is fully documented (see also Appendix I). The careful reader can see how the results of the needs analysis influenced the adoption of certain test formats (dictation, the integration of reading, listening and writing activities) and clarified the range of skills to be tested and the assessment criteria to be employed. The recent IELTS revision project has forsaken such needs analysis and instead intends to rely on post hoc 'expert' comment to judge the 'authenticity' and other aspects of the products of the test writing teams, who were not to be constrained by any a priori specifications (see Appendix V).

In the development of tests in the future the balance of attention must be paid to translating specifications into test realisations and validating the latter (in the ways to be discussed in Chapter Two), though the need to specify in advance of item writing the construct that is to be measured is ignored at one's peril. The emphasis, however, should be on test development and validation rather than on the analysis of needs for creating test specifications. For this reason the discussion of needs analysis has been limited, and instead the focus in Chapters Three and Four will be on test construction and in particular on examining a range of possible formats for testing language skills within a communicative framework.

The crucial stage in any test development occurs when the specification is translated into a test realisation. The test that results should exhibit the qualities of validity, efficiency and reliability which are examined next in Chapter Two. These qualities need to be determined both qualitatively a priori and empirically a posteriori.