Chapter Two

Basic Considerations in Test Design

The concepts of validity, reliability and efficiency affect all aspects of test design, irrespective of the prevailing linguistic paradigm. In Chapter One the relative importance of these concepts in recent approaches to language testing was reviewed. In this chapter the nature of these key concepts is examined in more detail. The status of the various types of validity, and how the concept of validity relates to those of efficiency and reliability are examined. Chapters Three and Four are more practically oriented; they examine the specification and realisation of the theoretical foundations discussed in the first two chapters.

2.1 The concept of validity

2.1.1 Construct validity

The concept of validity (does the test measure what it is intended to measure?) can be approached from a number of perspectives; the relationship between these is interpreted in a number of ways in the literature.

The most helpful exegesis regards construct validity as the superordinate concept embracing all other forms of validity. Anastasi (1982, p. 153) is of the opinion that: 'content, criterion-related and construct validation do not correspond to distinct or logically co-ordinate categories. On the contrary, construct validity is a comprehensive concept which includes the other types.'

Cronbach (1971, p. 463) commented that: 'Every time an educator asks "but what does the instrument really measure?" he is calling for information on construct validity.' Anastasi (1982, p. 144) defined it as:

> the extent to which the test may be said to measure a theoretical construct or trait .... Each construct is developed to explain and organise observed response consistencies. It derives from established inter-relationships among behavioral measures .... Focusing on a broader, more enduring and more abstract kind of behavioral description ... construct validation requires the gradual accumulation of information from a variety of sources. Any data throwing light on the nature of the trait under consideration and the conditions affecting its development and manifestations are grist for this validity mill.

She argued that the theoretical construct, trait or behaviour domain measured by any test can be defined in terms of the operations performed in establishing the validity of the test. She was careful to emphasise that the construct measured by a particular test (1982, p. 155): 'can be adequately defined only in the light of data gathered in the process of validating that test ... . It is only through the empirical investigation of the relationship of test scores to other external data that we can discover what a test measures.'

The view expressed below differs only insofar as external empirical data are seen as a necessary but not a sufficient condition for establishing the adequacy of a test for the purpose for which it was intended. Though there is a lack of an adequate theoretical framework for the construction of communicative tests, this does not absolve test constructors from trying to establish a priori construct validity for a test conceived within the communicative paradigm. A test should always be designed on a principled basis, however limited the underlying theory, and, wherever possible after its administration, statistical validation procedures should be applied to the results to determine how successful the test has been in measuring what it intended to measure.

In the past little attention has been accorded to the non-statistical aspects of construct validity. In the earlier psychometric-structuralist approach to language testing (see Section 1.2) the prevailing theoretical paradigm lent itself easily to testing discrete elements of the target language and little need was seen for much a priori deliberation on the match between theory and test. Additionally the empiricism and operationalism of those working in educational measurement made the idea of working with non-objective criteria unattractive. The notions of concurrent and predictive validity, more consistent with the principles of operationalism and the desire for an objective external criterion, took precedence.

Construct validity is viewed from a purely statistical perspective in much of the recent American literature (see Palmer *et al.* 1981; Bachman and Palmer, 1981a). It is seen principally as a matter of the a posteriori statistical validation of whether a test has measured a construct which has a reality independent of other constructs. The concern is much more with the a posteriori relationship between a test and the psychological abilities, traits and constructs it has measured than with what should have been elicited in the first place.

To establish the construct validity of a test statistically, it is necessary to show that it correlates highly with indices of behaviour that one might theoretically expect it to correlate with and also that it does not correlate significantly with variables that one would not expect it to correlate with. An interesting procedure for investigating this is the convergent and discriminant validation process first outlined by Campbell and Fiske (1959) and later used by Bachman and Palmer (1981b). The latter argue that the strong effect of test method that they discovered points to the necessity of employing a multi-trait multi-method matrix as a research paradigm in construct validation studies. They found that the application of confirmatory factor analysis to these data enabled them to quantify the effects of trait and method on the measurements of proficiency employed and provided a clearer picture of this proficiency than was available through other methods.

The experimental design of the multi-trait/multi-method matrix has been criticised (see Low, 1985) especially in relation to more direct tests of language proficiency, but nevertheless is deserving of further empirical investigation as so few studies have been reported, particularly from this side of the Atlantic. It is a potentially useful, additional measure for clarifying what it is that we have measured in a particular application of a test. The only difficulty in employing this technique is that to be effective a high degree of test reliability is essential as error variance is likely to confound the results.

In contrast to this emphasis on a posteriori statistical validation there is a body of opinion which holds that there is an equally important need for construct validation at the a priori stage of test design and implementation.

Cronbach (1971, p. 443) believes that: 'Construction of a test itself starts from a theory about behaviour or mental organisation derived from prior research that suggests the ground plan for the test.' Davies (1977, p. 63) argued in a similar vein: 'it is, after all, the theory on which all else rests; it is from there that the construct is set up and it is on the construct that validity, of the content and predictive kinds, is based.' Kelly (1978, p. 8) supported this view, commenting that: 'the systematic development of tests requires some theory, even an informal, inexplicit one, to guide the initial selection of item content and the division of the domain of interest into appropriate sub-areas.'

It would seem self-evident that the more fully we are able to describe the theoretical construct we are attempting to measure, at the a priori stage, the more meaningful might be the statistical procedures contributing to construct validation that can subsequently be applied to the results of the test. Statistical data do not in themselves generate conceptual labels. We can never escape from the need to provide clear statements concerning what is being measured, just as we are obliged to investigate how adequate a test is in operation, through available statistical procedures.

2.1.2 Content validity

Because we lack an adequate theory of language in use, a priori attempts to determine the construct validity of proficiency tests involve us in matters which relate more evidently to content validity. The more a test simulates the dimensions of observable performance and accords with what is known about that performance, the more likely it is to have content and construct validity. We can often only talk about the communicative construct in descriptive terms and, as a result, we become involved in questions of content relevance and content coverage. Thus, for Kelly (1978, p. 8) content validity seems 'an almost completely overlapping concept' with construct validity, and for Moller (1982b, p. 68): 'the distinction between construct and content validity in language testing is not always very marked, particularly for tests of general language proficiency.'

Given the restrictions on the time and resources available to those involved in test construction, especially for use in the classroom, it is often only feasible to focus on the a priori validation of test tasks. In these cases, particular attention must be paid to content validity in an attempt to ensure that the sample of activities to be included in a test is as representative of the target domain as is possible.

A primary purpose of many communicative tests is to provide a profile of the student's proficiency, indicating in broad terms the particular modes where deficiencies lie. Content validity is considered especially important for achieving this purpose as it is principally concerned with the extent to which the selection of test tasks is representative of the larger universe of tasks of which the test is assumed to be a sample (see Bachman and Palmer, 1981a).

Anastasi (1982, p. 131) defined content validity as: 'essentially the systematic examination of the test content to determine whether it covers a representative sample of the behaviour domain to be measured.' She (p. 132) provided a set of useful guidelines for establishing content validity:

1. 'the behaviour domain to be tested must be systematically analysed to make certain that all major aspects are covered by the test items, and in the correct proportions';
2. 'the domain under consideration should be fully described in advance, rather than being defined after the test has been prepared';
3. 'content validity depends on the relevance of the individual's test responses to the behaviour area under consideration, rather than on the apparent relevance of item content.'

The directness of fit and adequacy of the test sample is thus dependent on the quality of the description of the target language behaviour being tested.

J.B. Carroll (1961) pointed to the importance of and the difficulties involved in defining the area of language from which the sample is to be taken and the resultant problems this has for sampling. Moller (1982b, p. 37) also referred to the problems involved: 'In the case of a proficiency test, however, the test constructors themselves decide the "syllabus" and the universe of discourse to be sampled. The sampling becomes less satisfactory because of the extent and indeterminate nature of that universe.'

Establishing content validity is problematic, given the difficulty in characterising language proficiency with sufficient precision to ensure the representativeness of the sample of tasks included in a test. Additional threats to validity may arise out of attempts to operationalise real-life behaviour in a test especially where some sort of quantification is necessary either in the task or the method of assessment. These difficulties do not, however, absolve the test constructor from attempting to make tests as relevant in terms of content as is possible.

The procedure of designing a test from a skills specification (see the list drawn up by Munby, 1978 and its attempted implementation by B.J. Carroll, 1981b) may lead to variability in opinions as to what is being tested by specific items. There is a need to establish clear procedures that might reduce this variability.

Further, there is a need to look closely at test specifications to make sure that they describe adequately what ought to be tested. A close scrutiny of the specification for a proficiency test by experts in the field (or colleagues in the case of classroom achievement tests) and the relating of the specification to the test as it appears in its final form is essential (see Weir, 1983a). This would provide useful information as to what the test designer was intending to test and how successful the item writers had been in implementing the specification in the test realisation.

Mere inspection of the modules in the test, even by language and subject experts, does

not necessarily guarantee the identification of the processes actually used by candidates in taking them. In addition, it would be valuable to employ ethnographic procedures to establish the validity of items in a test.

A useful procedure is to have a small sample of the test population introspect on the internal processes that are taking place in their completion of the test items (see Aslanian, 1985; Cohen, 1985). This would provide a valuable check on experts' surface-level judgements on what was being tested and would contribute to the establishment of guidelines for the conduct of this type of methodological procedure in future investigations of test validity.

It is crucial for a test supposedly based on specified enabling skills to establish that it conforms to the specifications, especially if claims are made for these being representative of the domain in question. To the extent that the content is made explicit the concern also becomes one of face validity which Porter (1983) describes as perhaps the most contentious validity that might be invoked.

2.1.3 Face validity

Anastasi (1982, p. 136) pointed out that face validity:

> is not validity in the technical sense; it refers, not to what the test actually measures, but to what it appears superficially to measure. Face validity pertains to whether the test 'looks valid' to the examinees who take it, the administrative personnel who decide on its use, and other technically untrained observers. Fundamentally, the question of face validity concerns rapport and public relations.

Lado (1961), Davies (1965), E. Ingram (1977), Palmer (1981) and Bachman and Palmer (1981a) have all discounted the value of face validity. Bachman and Palmer (1981a, p. 55) argue as follows:

> Since there is no generally accepted procedure for determining whether or not a test demonstrates this characteristic, and since 'it is not an acceptable basis for interpretative inferences from test scores', we feel it has no place in the discussion of test validity.

If a test does not have face validity though, it may not be acceptable to the students taking it, or the teachers and receiving institutions who may make use of it. If the students do not accept it as valid, their adverse reaction to it may mean that they do not perform in a way which truly reflects their ability. Anastasi (1982, p. 136) took a similar line:

> Certainly if test content appears irrelevant, inappropriate, silly or childish, the result will be poor co-operation, regardless of the actual validity of the test. Especially in adult testing, it is not sufficient for a test to be objectively valid. It also needs face validity to function effectively in practical situations.

The usual empirical caveat of course applies (Anastasi, 1982, p. 136): 'To be sure, face validity should never be regarded as a substitute for objectively determined validity .... The validity of the test in its final form should always be directly checked.'

Stevenson (1985b) expresses a similar concern that construct and content validities should not be sacrificed at the altar of an increased lay acceptance of non-technical face validity.

2.1.4 Washback validity

The difficulties of precisely determining what it is that needs to be measured perhaps argues for a greater concern with what has recently been termed 'washback validity' (Morrow, 1986) or more commonly (Porter, 1983 and Weir, 1983a) the washback of the test on the teaching and learning that precedes it.

Given that language teachers operating in a communicative framework normally attempt to equip students with skills that are judged to be relevant to present or future needs, and to the extent that tests are designed to reflect these, the closer the relationship between the test and the teaching that precedes it, the more the test is likely to have construct validity.

In other circumstances the tail may wag the dog in that a communicative approach to language teaching is more likely to be adopted when the test at the end of a course of instruction is itself communicative. A test can be a very powerful instrument for effecting change in the language curriculum as recent developments in language tests in Sri Lanka have shown.

A suitable criterion for judging communicative tests in the future might well be the degree to which they satisfy students, teachers and future users of test results, as judged by some systematic attempt to gather quantifiable data on the perceived validity of the test. If the test passes the first a priori validity hurdle it is then worthwhile establishing its validity against external criteria, through confirmatory a posteriori statistical analysis. If the first stage, with its emphasis on construct, content, face and washback validities, is bypassed then we should not be too surprised if the type of test available for external validation procedures does not suit the purpose for which it was intended.

For construct, content, face and washback validity, knowing what the test is measuring is crucial. There is a further type of validity which we might term criterion-related validity where knowing exactly what a test measures is not so crucial.

2.1.5 Criterion-related validity

This is a predominantly quantitative and a posteriori concept, concerned with the extent to which test scores correlate with a suitable external criterion of performance: what Ingram (1977, p. 18) termed 'pragmatic validity'. Criterion-related validity divides into two types (see Davies, 1977), concurrent validity, where the test scores are correlated with another measure of performance, usually an older established test, taken at the same time (see Kelly, 1978; Davies, 1983) and predictive validity, where test scores are correlated with some future criterion of performance (see Bachman and Palmer, 1981a).

For many authorities, external validation based on data is *always* superior to the 'armchair speculation of content validity'. Davies (1983, p. 1) has argued forcefully that external validation based on data is always to be preferred: 'The external criterion, however hard to find and however difficult to operationalise and quantify remains the best evidence of a test's validity. All other evidence, including reliability and the internal validities is essentially circular.' And he quotes Anastasi on the need for independently gathered external data: 'Internal analysis of the test, through item-test correlations, factorial analysis of test items, etc., is never an adequate substitute for external validation.'

Though this concept of criterion-related validity is more in keeping with the demands of an empiricist-operationalist approach, the problem remains that a test can be valid in this way without our necessarily knowing what the test is measuring, i.e., it tells us nothing directly about its construct validity. Morrow (1979, p. 147) drew attention to the essential circularity of employing these types of validity in support of a test:

> Starting from a certain set of assumptions about the nature of language and language learning will lead to language tests which are perfectly valid in terms of these assumptions but whose value must inevitably be called into question if the basic assumptions themselves are challenged.

For Jakobovits (1970, p. 75) the very possibility of being able to construct even one communicative test appeared problematic: 'the question of what it is to know a language is not well understood and, consequently, the language proficiency tests now available and universally used are inadequate because they attempt to measure something that has not been well-defined.'

Even if it were possible to construct a valid communicative test there would still be problems in establishing sufficiently valid criterion measures against which to correlate it. Hawkey (1982, p. 153) felt this to be particularly problematic for tests conceived within a communicative paradigm: 'At this developmental stage in communicative testing, other tests available as criteria for concurrent validation are likely to be less integrative/communicative in construct and format and thus not valid as references for direct comparison.'

There is a distinct danger that one might be forced to place one's faith in a criterion measure which may in itself not be a valid measure of the construct in question. One cannot claim that a test has criterion-related validity because it correlates highly with another test, if the other test itself does not measure the criterion in question.

It seems pointless to validate a test conceived within the communicative paradigm against tests resulting from earlier paradigms if they are premised on s'/ch different constructs. Similarly if equivalent but less efficient tests are not available against which to correlate, other external non-test criteria might need to be established.

Establishing these non-test criteria for validating communicative tests could well be problematic. Even if one had faith in the external criterion selected, for example, a sample of real life behaviour, the quantification process which it might be necessary to subject this behaviour to in order for it to become operational might negate its earlier direct validity.

Though caution is advocated in the interpretation of these criterion-related validity measures, they are still considered to be potentially useful concepts. For example, one might be very wary of tests that produced results seriously at variance with those of other tests measuring the same trait, especially if the latter had been found to have construct validity.

It is particularly important to try to establish criterion-related validity for a test through empirical monitoring whenever the candidates' futures may be affected by its results. For example, given the variety of language qualifications currently acceptable as evidence of language proficiency for entry into tertiary-level study in Britain (see Weir, 1983a), there is some cause for concern about the equivalence of such a broad spectrum of tests. Where is the empirical evidence for the equivalence of one entry qualification with another?

In the case of predictive validity, it may be that in certain circumstances the predictive power of the test is all that is of interest. If all one wants is to make certain predictions about future performance on the basis of the test results, this might entail a radically different test from that where the interest is in providing information to allow effective remedial action to be taken. If predictions made on the basis of the test are reasonably accurate then the nature of the test items and their content might not be so important.

Incidentally, both validity and reliability estimates based on correlational data must be treated with caution. A high correlation may indicate the measurement of two different attributes which are themselves quite highly correlated among the population of examinees. On the other hand, a low correlation may indicate that two quite different attributes are being measured or may merely reflect a high level of error variance in one or both of the tests.

2.1.6 How should a test be known?

Most GCSE examinations and existing language proficiency examinations, e.g., the University of Cambridge Local Examinations Syndicate (UCLES) Certificate of Proficiency in English (CPE) and the Joint Matriculation Board (JMB) Test in English (Overseas), because of their public, operational nature, are not over interested in concurrent or predictive validity, whereas, as Davies (1982) has pointed out, these are matters of major concern for most standardised, closed EFL tests. Correlating the results of one year's examination with other examinations or against some future criterion is perhaps viewed as a pointless exercise when a new set of examinations is already in preparation for the following year and results already issued for current candidates.

Only closed tests, such as the Associated Examining Board's Test in English for Educational Purposes, the TEEP test (see Appendix I), or UCLES and the British Council's English Language Testing Service's (ELTS) battery (see Appendix V), feel obliged to concern themselves with a posteriori validation procedures. Open examinations which are held annually tend to rely more heavily on construct (non-statistical), content and face validity.

In situations where the test is to have a diagnostic function a high degree of explicitness at the a priori stage of test construction is felt to be necessary. This is particularly so where the aim is to provide meaningful statements on a candidate's performance which would be of use to those providing remedial support for candidates with known difficulties.

If the concern is to collect appropriate information on a candidate's performance for the purposes of profile reporting rather than to establish a test's predictive validity, then there is more obligation to improve the content/construct validity of the test by identifying, prior to test construction, appropriate communicative tasks which it should include. This a priori validation is essentially a first, though crucial, step in the total validation process of a test.

Having made rigorous attempts at an a priori stage to make the test as valid as possible, there is then a need to establish the validity of the test against external criteria. If the first stage with its emphasis on content validity is bypassed, then the type of test available for

external validation procedures would not, in all likelihood, suit the purpose for which the test is intended.

To illustrate the recent awakening of interest in a priori validation of tests it might be useful to take a concrete example of the construction of a test for a particular purpose. Let us assume the task is to construct a proficiency test in English for Academic Purposes (EAP) which can also provide through profiling some diagnostic information on the language-related study skills candidates are weak in.

A test of discrete grammatical items constructed for this purpose might be found to correlate highly with an external criterion, e.g., another established test concurrently administered or a measure taken at a later date, such as final academic grades. That is, it might possess criterion-related validity. It would, however, be of less value to those providing remedial English language support, who, rather than a single score, require information about the particular study modes in which a student has difficulty operating, i.e., they might be better served by a test exhibiting construct, content and face validity. One would not be able to allocate students effectively to remedial language programmes on the basis of performance in a discrete-point structuralist test lacking these validities.

The a priori validation of an EAP proficiency test with diagnostic potential would seem to demand that we test integrated macro-skills rather than micro-elements in isolation. If the aim is to test the communicative competence of overseas students in an EAP setting, it is doubtful whether tests of linguistic competence alone are appropriate because the constructs for such tests are necessarily based on discrete linguistic levels, not on integrative work samples. Since the essence of communication is an ability to combine discrete linguistic elements in a particular context it seems essential that this ability should be assessed by tests of integrated skills rather than by tests of discrete linguistic levels in isolation.

The content of an EAP proficiency test based on work samples from the target situation would be qualitatively different from the content of a test of linguistic competence based upon discrete linguistic items. In the case of the EAP proficiency test which aims at assessing communicative competence the main justification for item selection would be a careful sampling of the communicative tasks required of students in English-medium study. In the case of a test of linguistic competence a test may be considered valid if its content is based on an adequate sample of 'typical' discrete linguistic elements.

According to Canale and Swain (1980, p. 34) communicative testing: 'must be devoted not only to what the learner knows about the second language and about how to use it (competence) but also to what extent the learner is able to actually demonstrate this knowledge in a meaningful communicative situation.'

The proficiency tester today is influenced by what Moller (1981b) has described as the sociolinguistic—communicative paradigm. The nature of communicative testing was discussed in Section 1.4 above. Briefly a test within this communicative paradigm might be expected to exhibit the following features:

- There would be an emphasis on interaction between participants, and the resultant intersubjectivity would determine how the encounter evolves and ends.
- The form and content of the language produced would be to some extent unpredictable.
- It would be purposive in the sense of fulfilling some communicative function.

- It would employ domain-relevant texts and authentic tasks.
- Abilities would be assessed within meaningful and developing contexts and a profile of performance on these made available.
- Where deemed appropriate and feasible, there might be an integration of the four skills of reading, listening, speaking and writing.
- The appropriateness of language used for the expression of functional meaning would have high importance.
- It would use direct testing methods, with tasks reflecting realistic discourse processing.
- The assessment of productive abilities would most probably be qualitative rather than quantitative, involving the use of rating scales relating to categories of performance.

Thus a good deal more attention will have to be paid to content and face validity than was the case within previous orthodoxies. However, given the rudimentary state of the art in communicative approaches to language testing, some authorities still feel it would be prudent to retain a number of components which sample major linguistic categories, for, as Moller (1981b, p. 44) argued:

> It is clear that communicative testing does test certain aspects of proficiency. But it is important to be aware that testing language proficiency does not amount just to communicative testing. Communicative language performance is clearly an element in, or a dimension of, language proficiency. But language competence is also an important dimension of language proficiency and cannot be ignored. It will also have to be tested in one or more of the many ways that have been researched during the past thirty years. Ignoring this dimension is as serious an omission as ignoring the re-awakening of traditional language testing in a communicative setting.

The revision of the British Council/UCLES ELTS Test 1986-89, which has resulted in the new IELTS battery (see Appendix V), had originally planned to include a test of lexis and grammar in the General Component. In the earlier trials of the TEEP test 1979-82 a discrete item multiple-choice test of grammar had been included in the trial battery and proved to be a robust and valid indicator of General Language Proficiency. The TEEP research (Weir 1983a) indicated clearly, however, that the grammar component added no additional information to the picture of a candidate's language ability already available from the more communicative, use-based components. For this reason it was dropped from the battery. For similar reasons the tests of lexis and grammar have been dropped from the IELTS battery.

So far we have concentrated on examining ways of improving the validity of tests and neglected the crucial fact that unless a test is reliable it cannot be valid. The need for reliability in order to guarantee the validity of our tests is the next issue we address.

2.2 The concept of reliability

A fundamental criterion against which any language test has to be judged is its reliability (see Anastasi, 1982; Guilford, 1965). The concern here is with how far can we depend on the results that a test produces or, in other words, could the results be produced consistently.

Three aspects of reliability are usually taken into account. The first concerns the consistency of scoring among different markers, e.g., when marking a test of written expression. The degree of inter-marker reliability is established by correlating the scores obtained by candidates from marker A with those from marker B. The consistency of each individual marker (intra-marker reliability) is established by getting them to remark a selection of scripts at a later date and correlating the marks given on the two occasions. (See Anastasi, 1982 for a clear and accessible introduction to the necessary statistics. Also of use are Crocker, 1981 and more recently Woods *et al.*, 1986.)

The concern of the tester is how to enhance the agreement between markers by establishing, and maintaining adherence to, explicit guidelines for the conduct of this marking. The criteria of assessment need to be established and agreed upon and then markers need to be trained in the application of these criteria through rigorous standardisation procedures (see Murphy, 1979). During the marking of scripts there needs to be a degree of cross-checking to ensure that agreed standards are being maintained.

It is also considered necessary to try and ensure that relevant sub-tests are internally consistent in the sense that all items in a sub-test are judged to be measuring the same attribute. The Kuder-Richardson formulae for estimating this internal consistency are readily available in most statistics manuals (see Anastasi, 1982, pp. 114—6).

The third aspect of reliability is that of parallel-forms reliability, the requirements of which have to be borne in mind when future alternative forms of a test have to be devised. This is often very difficult to achieve for both theoretical and practical reasons. To achieve it, two alternative versions of a test need to be produced which are in effect clones of each other. The reliability of the versions is directly proportional to the similarity of the results obtained when administered to the same test population. Less frequently reliability is checked by the test-retest method where the same test is readministered to the same sample population after a short intervening period of time.

The concept of realiability is particularly important when considering language tests within the communicative paradigm (see Porter, 1983). For as Davies (1965, p. 14) stressed: 'reliability is the first essential for any test; but for certain kinds of language test may be very difficult to achieve.'

2.3 Validity and reliability — an inevitable tension?

Given the normal limitations affecting test development (especially of achievement tests in the classroom), concern usually centres on validation at the test construction stage and only to a lesser extent with a posteriori validation at the performance stage. The resources to do large-scale concurrent and predictive validity studies, such as conducted by Moller (1982b) and by the Institute of Applied Language Studies at the University of Edinburgh, on the ELTS battery, are not normally available.

The concern is often by necessity with content, construct and face validity though the predictive and concurrent validity of tests should always be examined as far as circumstances allow. Validation might prove to be a sterile endeavour, however, unless care has also been taken over test reliability.

The problem is that while one can have test reliability without test validity a test can only be valid if it is also reliable. There is thus sometimes said to be a reliability—validity tension (see Guilford, 1965 and Davies, 1978). This tension exists in the sense that it is sometimes essential to sacrifice a degree of reliability in order to enhance validity. If, however, validity is lost to increase reliability we finish up with a test which is a reliable measure of something other than what we wish to measure. The two concepts are, in certain circumstances, mutually exclusive, but if a choice has to be made, validity 'after all, is the more important', (see Guilford, 1965, p. 481).

Rea (1978) argued that simply because tests which assess language as communication cannot automatically claim high standards of reliability in the same way that discrete-item tests are able to, this should not be accepted as a justification for continued reliance on highly reliable measures having very suspect validity. Rather, we should first be attempting to obtain more reliable measures of communicative abilities. This seems a less extreme and more sensible position than that adopted by Morrow (1979, p. 151), who argued polemically that: 'Reliability, while clearly important, will be subordinate to face validity. Spurious objectivity will no longer be a prime consideration.'

Rea's viewpoint was shared by Read (1981a, p. x—xi), who reported that a recurring theme at the April 1980 RELC Seminar on 'Evaluation and Measurement of Language Competence and Performance' was that: 'subjective judgements are indispensable if we are to develop testing procedures that validly reflect our current understanding of the nature of language proficiency and our contemporary goals in language teaching.'

Read went on to emphasise that: 'this does not mean a return to the old pre-scientific approach. It is generally accepted that a substantial, verifiable level of reliability must also be attained, if test results are to have any meaning.' Moller adopted a similar approach (1981a, p. 67):

> While it is understood that a valid test must be reliable, it would seem that in such a highly complex and personal behaviour as using a language other than one's mother tongue, validity could be claimed for measures that might have a lower than normally acceptable level of reliability.

He argued that, although reliability is something we should always try to achieve in our tests, 'it may not always be the prime consideration' and offers a possible compromise position (p. 67):

> In constructing test batteries that contain different types of task, for example, certain of the sub-tests may be required to exhibit a high degree of reliability. Other sub-tests, particularly tests of communicative use, may quite properly exhibit lower reliability without adversely affecting the overall validity of the battery.

Hawkey (1982, p. 149) commented in a similar vein:

> the reliability of a test cannot be ignored without a harmful effect on the validity of the instrument. But it is likely that, if the construct validity of communicative tests is to be ensured, the reliability question is going to have to be accepted as subordinate, though worked at fairly hard by item analysis and correlational operations.

Validity is important also because it is related to the way in which test performance

levels are defined. Houston (1983) describes the difference between norm- and criterion-referenced methods of defining levels and discusses some of the difficulties of specifying appropriate performance criteria when the latter method is chosen. Popham (1978, p. 2) provided the following functional definitions of these approaches:

> a criterion-referenced test is designed to produce a clear description of what an examinee's performance on the test actually means. Rather than interpreting an examinee's test performance in relationship to the performance of others as is the case with many traditional tests, a good criterion-referenced test yields a better picture of just what it is that the examinee can or cannot do.

Davies (1978, p. 158) made the connection with language testing and expressed certain reservations about criterion-referenced tests:

> there are difficulties in using criterion-referenced tests for language: there is no finite inventory of learning points or items; there are very many behavioural objectives; there are variable (or no) external criteria of success, fluency, intelligibility, etc; there is no obvious way of establishing adequate knowledge, of saying how much of a language is enough.

Thus some of the difficulties referred to later by Houston (1983) are put in a language testing context. Clearly, criterion-referencing of performance levels is possible only to the extent that the test has a high degree of content validity.

2.4 Test efficiency

A valid and reliable test is of little use if it does not prove to be a practical one. This involves questions of economy, ease of administration, scoring, and interpretation of results. The longer it takes to construct, administer and score, and the more skilled personnel and equipment that are involved, the higher the costs are likely to be.

The duration of the test may affect its successful operation in other ways, e.g., a fatigue effect on the candidates, administrative factors such as staff to invigilate and the availability of rooms in which to sit the examination; all have to be taken into consideration. It is thus highly desirable to make the test as short as possible, consistent with the need to meet the validity and reliability criteria referred to above. If the aim is to provide as full a profile of the student's abilities as is possible then there is obviously a danger of conflict, for although hard-pressed administrators seem to want a single overall grade, remedial language teachers would prefer as much information as possible (see Moller, 1977; Alderson and Hughes, 1981; and Porter, 1983).

To provide profiles rather than standard scores, each part of the profile will need to reach an acceptable degree of reliability. To achieve satisfactory reliability, communicative tests may have to be longer and have multiple scoring. The difficulties in ensuring that the test contains a representative sample of tasks may also serve to lengthen it. To enhance validity by catering for specific needs and profiling, more tests will be needed, thus further raising the per-capita costs as compared to those of single general tests available for large populations.

Efficiency in the sense of financial viability, may prove the real stumbling block in the way of the development of communicative tests. Tests of this type are difficult and time consuming to construct, require more resources to administer, demand careful training and standardisation of examiners and are more complex and costly to mark and report results on. The increased per-capita cost of using communicative tests in large-scale testing operations may severely restrict their use.

However problematic, there is clearly an imperative need to try and develop test formats and evaluation criteria that provide the best overall balance among reliability, validity and efficiency in the assessment of communicative skills. In the survey of developments in language testing in Chapter One we noted that the pendulum of change had swung several times and differing emphases had been given in test design and implementation to the demands of reliability, validity and efficiency. In this chapter these concepts have been examined from a deeper, theoretical perspective.

In Chapter Three we return to more practical concerns and the stages in the development of a test are briefly outlined to give an idea of the processes that are normally followed in the design and implementation of a language test. This is followed in Chapter Four by an examination of a range of formats available for testing language skills within the communicative paradigm and an assessment is made of their advantages and disadvantages in the light of the discussion in Chapters One and Two.

Chapter Four is intended to be of immediate practical use to the reader faced with the problem of selecting appropriate test formats. It outlines what might be the best choices given the uncertain state of the art in communicative testing and a desire, nevertheless, to make a test as communicative as possible within the constraints imposed by considerations of practicality and reliability.