

Chapter 1

TESTING AND EVALUATION IN LANGUAGE TEACHING

1.1 Aims and scope of this course

The aims of this course are to encourage reflective language teachers, and others interested in the teaching and learning of languages in classroom settings, to explore and develop the educational relationships that hold -- and could hold -- between testing, evaluation and teaching. Taking the language classroom as its point of departure and return, the course focuses on concerns about language learning and language teaching. It emphasises the search for evidence to support or undermine claims about successful learning, learners' persistent difficulties, or effects of particular activities and practices on learners and learning. The course reflects the writer's conviction that an active involvement in some of the procedures and findings of language testing and language programme evaluation can help intending and practising teachers to further understand and enhance what is taking place in and around language classes.

The main justification for writing the course is that the language teaching profession has lacked a basic introductory account of language testing and evaluation. This state of affairs seems strange, given the persistent occurrence of "language testing and evaluation" in the titles of courses, conferences and committees or working groups in language education. But perhaps my comment should be reversed: we could argue instead that it is the persistence of efforts by planners to link language testing and language programme evaluation together that should call for critical scrutiny. Certainly, to date, these attempts do not appear to have been conspicuously successful. A major reason for this, I suggest, is that language testing has been the more established field within applied linguistics and language education, and that "evaluation" is too often appended to "language testing" as something of a fashionable afterthought. (In a major conference on "Language Testing and Evaluation" held in

Singapore in 1990, "evaluation" featured as the fourth day of four.) Whether desirable or not, a synthesis of these areas within language teaching has yet to be achieved, although the work of some applied linguists who publish both in language testing and in evaluation is starting to contribute towards this end. A basic introductory text obviously cannot suffice to integrate the two areas, but it can point to some of the relations that people have seen or forged between them, and thereby encourage greater curricular coherence.

If one recognises that learning conditions and learning outcomes are important concerns for teachers, the potential contribution of language testing and evaluation procedures to teaching might easily be presentable as commonsensical and self-evident. To offer any such account, however, would obscure some important reservations that can arise among educators and others regarding both testing and evaluation. Some educators still seek to distance themselves from these aspects of professional life, and others find that practices in testing and evaluation fall short of, or even conflict with, what is educationally desirable. Although the nature and extent of these sorts of reactions vary markedly across communities and circumstances, general concerns about the purposes, values and practices that underlie language testing and language programme evaluation need to be identified and addressed. A questioning stance also has a great deal to commend it as a guide to improving professional practice.

Although the course is concerned with how issues in language testing and evaluation affect classroom practices, this carries no implication of narrowness in scope. On the contrary, discussions on the course will come to grips with some of the wider challenges and concerns over authority in language, language teaching and education that are prominent in contemporary educational debates. It will be seen, for example, that community notions about language standards (and standard languages) are relevant to the goals of language teaching, to views about the purpose and desired nature of language tests and examinations, and to possible bases for evaluating the success of language programmes in their educational and cultural settings.

The author's general view is that testing and evaluation are of considerable potential value in the pursuit of language education, even though they can easily be pursued in ways that may become too narrow and reductive to realise this value effectively. It follows, on this view, that

critics could more usefully target bad practices than all practices in language testing and in language programme evaluation. Whatever views and experiences readers bring to these issues, I hope that they will find this introductory text useful for their purposes in investigating the field.

Activity 1.1 - Point of Departure

- What are your own main learning goals for this course at this stage?
- What reservations, if any, do you have about language testing and examining, and about evaluations of language teaching, as these are carried out in teaching-learning situations known to you?
- What reservations, if any, have you heard other people express in these areas?
- Assuming that you were to meet an educator from a very different and unfamiliar education system, what could you most usefully tell this person about language testing and evaluation issues within the local context known to you? Also, what questions would you ask about language testing and evaluation in the unfamiliar context?

1.2 Testing, evaluation and teaching: terms and contexts

Language testing involves the assessment of some or all aspects of the language ability of individuals in some context (not necessarily that of a language class) and for some set of purposes (not necessarily common to all parties). Language ability will be discussed in chapter 3, and the specific theme of classroom language testing and its purposes will be pursued in chapter 5. "Testing" is sometimes used almost interchangeably with "assessment" and in this spirit is taken here as a broad cover term for both formal and informal assessment procedures. In such a sensitive area, even such preliminary choices of terminology can evoke particular attitudes and ideologies for different people, so my own choices call for comment.

For some speakers and writers, "testing" is used more narrowly to denote only those formal modes of assessment that are officially scheduled, with clearly delimited time on task and strict limitations on available guidance. The somewhat provocative title of Hill and Parry's

edited volume (1994), *From testing to assessment*, proposes a shift away from formal examinations and tests towards more and perhaps exclusive use of continuous assessment procedures. A related distinction in some people's usage seems to be that "testing", more than "assessment", evokes ideas of the measurement or estimation of abilities, notions that are sometimes represented as unduly restrictive. I do not myself see "assessment" as free of such associations either, even if the term lends itself well to indirect and covert forms of testing. It seems to me that the same issues of purpose, method and justification will need to be faced irrespective of one's choices between "testing" and "assessment" as general terms.

Activity 1.2 - What's in a name?

- How do you define "testing"? And "assessment"?
- What phrases come to your mind that include the words "test", "tests" or "testing"? What phrases do you associate with "assessment"?

There are no generally "right answers", of course, but your responses should afford an additional basis for comparison when you read on.

In this book, "testing" is not confined to formal modes of "assessment", and neither term is seen as free of association with exercises of power or responsibility on the part of teachers and examiners. My choice of "testing" as the general term in the title, and in much of the text, serves in part to emphasise continuity, rather than contrast, between formal and informal assessment modes. Though not motivated by any other conscious ideological preference, the choice also reflects my judgement of collocational probabilities. Whereas "language testing and evaluation" is quite common, I have yet to see a course or conference title exploring "language assessment and evaluation". I shall still use "assessment" in familiar collocations such as "continuous assessment" and "self-assessment", and elsewhere to reflect usage on the part of other authors I discuss.

"Evaluation" is also used in a variety of senses and contexts. It is always worth asking "evaluation of what, by whom, and for what purposes?" Our more specific concerns in this book are with the

evaluation of language programmes (introduced in chapter 4), especially by those involved in teaching them. A language programme evaluation sets out to establish the merits, limitations, and overall effectiveness and impact of a curriculum as it is actually realised in teaching and learning experiences. I use "language programme" more or less interchangeably with "language curriculum" to imply not only the stated goals but the actual experiences that make up a complete course of language studies. (The British spelling distinguishes this sense from the planned step-by-step explicitness of a computer program; the American spelling of language "program" is adopted only when I cite authors who use this spelling.) Language programme evaluations make use of many different forms of information, of which test results are just one instance. Among other characteristic evaluation procedures are classroom observation, document analysis, questionnaire administration to teachers and learners, group discussions and individual interviews with teachers and learners.

To avoid taking too much for granted in our discussions of programme evaluation, we shall also look into how language performance or language ability may be "evaluated" by different groups of people outside teaching contexts (chapter 2). This theme is relevant both to an understanding of language practices among different discourse communities and to communication between professional and other groups concerned with language teaching. These are important matters as teachers' and learners' goals will normally take account of the language values and expectations of professional and social groups beyond the language classroom itself, and evaluations of language programmes would ignore wider expectations and beliefs at their peril. An incidental point is that we should not expect to find terms like "assessment" and "evaluation" clearly differentiated in non-specialist discussions of language use and language standards.

Language teaching is not the only context in which language use is evaluated or language abilities are tested, but it is the context or set of contexts for which language testing and language programme evaluation are sometimes brought together. In discussions of language teaching, I believe it is convenient to follow the practice of Nunan (1992) in distinguishing "program evaluation" from "student assessment", and to take assessment (or testing) data as one source of information among others for programme evaluators to consider. The key theme that links testing and evaluation in the context of language teaching is that of

decision-making. While I shall not attempt a formal definition of teaching, "informed decision-making to promote learning" might well form part of any such definition. Information about language learner performance, obtainable through various kinds of testing, is clearly relevant in principle to decisions about what to teach or reteach as a course takes place, or about what to emphasise more or less in future versions of a course. Decision-making is also critical in accounts of language programme evaluation, which is distinguishable from other studies of language teaching precisely because it is oriented towards decisions about current or future practice.

Attempts to relate testing, evaluation and teaching The importance of decision-making is attested in a number of discussions that seek to relate language testing to evaluation, whether prominently as in Bachman (1990) or incidentally as in Henning (1987). Bachman (1990: 22-24) discusses the relationship between evaluation, measurement and testing along the following lines. All tests involve measurement, but not all measurement involves testing. Some evaluation also involves measurement, both in tests and in other forms of measurement, while some does not. Conversely, of course, not all tests and other measurements are used for evaluation. Bachman asserts that "It is only when the results of tests are used as a basis for making a decision that evaluation is involved" (Bachman, 1990: 22-23). For our purposes in this book, this comment would need to be further specified as a decision about some aspect of a language teaching programme. Test results are obviously also used when decisions are made about student placement, progress, final grade or certification, and so forth, but such decisions are not invariably taken up in the evaluation of the teaching programme itself.

Recent histories of language testing and evaluation These histories are quite distinct, and we can only touch upon them here. Language testing has long been an important area in applied linguistics, partly because constructs such as language proficiency have to be made explicit if they are to serve as models for test design and validation purposes. Validation involves ascertaining whether a test is effectively measuring what it was designed to measure (see also section 1.4). On the other hand, some of the specialised concerns of language testers (e.g. with aspects of measurement theory) have tended to keep them apart from other developments or

discussions in applied linguistics and language education. There seems to be a recent tendency to assert that language testing has more or less come of age, but how far this involves real progress is debatable (see further reading).

Evaluation has its own long tradition in education, but has until recently been substantially neglected in the context of English language teaching, at least according to the picture drawn by Beretta (1992). Beretta's discussion goes on to trace a current development of serious evaluation studies in applied linguistics. Murphy (1985) gives a picture of predominant neglect at the time, but more optimistically points out that a lot of teacher activity involves evaluation under other guises. Nunan (1992), quoting Brown (1989), draws attention to the idea that concerns over "needs analysis" in teaching English for specific purposes are not so far removed from questions of curriculum evaluation.

Evaluation as research? One other question to note is whether evaluation is a form of research or another kind of activity. The answer may not matter much in itself, but the different value systems involved in the discussion can be important. Mackay (1991) considers the two to be separate. Because of the way in which evaluations are commissioned, the questions are not decided by researchers themselves in the pursuit of disinterested knowledge. Mackay's view could, however, easily be challenged as a naive or idealised impression of what other research is like, and how it is commissioned. His account may also underestimate the capacity of evaluators to redefine their own briefs. In contrast, Nunan (1992: 193 and 196) takes evaluation to be a form of research. Although I have much sympathy with this view, it is not clear, at least to me, how it could be reconciled with some of Nunan's later comments on what counts as a researchable question. Elsewhere, Nunan views questions about what should be done as falling outside the scope of research questions (Nunan 1992:213). The evaluation studies in Alderson and Beretta (1992) obviously use research methodologies and habits of thinking. These constitute a compelling case for adopting research perspectives in carrying out programme evaluations. In the end, one's answer to this question will depend on an individual's personal philosophy, view of research and knowledge, and on the importance and value attached to studies carried out by practitioners (such as language teachers) rather than outside observers.

1.3 Controversies and constraints

Although language testing has long been a familiar practice in many educational settings, and explicit procedures for language programme evaluation have also become part of professional life for many language teachers, both areas of activity can give rise to controversy. Testing can be threatening to some test takers and others (including some parents and teachers), and the educational role of formal testing in particular is often contested. Reactions to concerns about language testing range from those who would do away with formal assessment altogether, e.g. Hill and Parry (1994), to those who believe that testing of both formal and informal kinds is here to stay and can be made to work for educational ends if we improve both testing practices and communication with others. An example of the latter position is Alderson (1991), who adds that testing is too important to be left to testers. Language programme evaluation is also much too important to be left to specialist evaluators. While the principle of programme evaluation seems unlikely to be widely challenged, actual practices in this area can easily be experienced as threatening to teachers, especially if they are linked to procedures for staff appraisal.

A common element in concerns over language testing and evaluation is the exercise of authority and power that they involve. Both forms of activity are liable to be associated with questions of teacher accountability to external forces. This association is sometimes accepted as a fact of life, but it can also give rise to sentiments of resistance or subversion, or to the nominal compliance of cynical alienation ("going through the motions"). Negative reactions become particularly likely, and understandable, where teachers feel that their educational experience, motivation and judgement are misunderstood, undervalued or conveniently disregarded by others with less understanding but greater power. At worst, testing and evaluation can both be perceived as imposed administrative requirements that lack educational warrant. In both areas, teachers who do not identify with the goals of an activity and who feel powerless to affect what is done and how it is done are unlikely to be convinced by assurances about educational benefits. This means that there are considerable challenges for effective educational management, including genuinely participatory consultations over goals and procedures, in the areas of language testing and programme evaluation.

Activity 1.3 - Where do you stand?

- How important are tests and examinations in the education system best known to you?
- What is your position on this situation? Are tests / examinations important enough? Are they too important? Are the tests / examinations in use appropriate for the educational goals that teachers value? What are these goals?
- Are language teaching programmes evaluated in a systematic way? Are they evaluated at all? If so, who carries out the evaluation? Do teachers have a role, and is this role a suitable one? Is evaluation also linked to staff appraisal, and should this be the case? Why / why not?

Explanations and invitations to discussion, on the part of educational managers, will need to be both frank and focused if they are to achieve their purposes. Some aspects of practice may well be non-negotiable, such as the fact that a school is preparing its students for national examinations. At worst, therefore, well-meaning but vague invitations to teachers to discuss the educational pros and cons of the examination system are easily viewed with further cynicism. ("Nothing's going to change, so why pretend to talk about it?") Discussions that focus upon actual teaching practices and opportunities for educational enrichment within an examinable course context will be more likely to help teachers engage seriously and constructively with issues on which they know they have an influence.

This is not to say, however, that current practices over such matters as examinations cannot be constructively challenged: that would be an unduly bleak picture in many settings. The nature of an examination system can change substantially over time in the light of wider discussions about the aims and priorities of the curriculum and the effects of existing examinations on the achievement or the thwarting of such aims. The long-term challenge for educational managers, then, is to involve teachers actively in giving shape to the educational goals of a curriculum, working within constraints of time and syllabus, but without regarding these as wholly immutable. Among other things, this requires honesty and explicitness on the part of managers about goals that they are pursuing

and constraints under which they themselves operate. Their concerns may often include limited budgets, or a need to convince other administrative bodies that a programme is giving value for money, or to satisfy public disquiet over standards being achieved. Both language testing and language programme evaluation are carried out for a variety of purposes, not all of them directly addressing the enhancement of learning, and these purposes also need to be clearly explained to teachers if full and active cooperation is to be won.

A need for clear explanations of principle and practice is not, of course, confined to educational management, but also applies to teachers themselves. In considering questions of administrative power and professional responsibility, we must recognise that teachers are the most obvious authority figures in their own classes and that they hold considerable power when it comes to the nature of the curriculum that students experience. This implies a responsibility to explain the curriculum both to learners and to others with a legitimate interest in what is going on in language classes. Through such openness, constraints can be better appreciated, and controversies may either be reduced or productively pursued in the interests of learning.

1.4 Fundamentals and directions

As it soon becomes apparent that language testing and evaluation are areas in which questions of value as well as issues of fact assume critical importance, I shall first briefly summarise my own stance so that readers can allow for it. The position taken throughout this course is that wider questions of educational accountability are better addressed through resolute individual and institutional professionalism than through strategies of avoidance. I also believe that an articulate professionalism offers the best way to acknowledge, and to work to reduce, the misunderstandings that easily arise among teachers, administrators, outside "experts", other professional and lay communities, parents, and learners. On this view, fully professional attitudes towards questions of language testing and evaluation will require an assumption, not an abdication, of responsibility for judgements on the part of teachers.

Much contemporary writing on language teaching assumes the importance of contributions made by reflective teachers for whom curriculum renewal and teacher development are intimately associated.

The present course is no exception, and will advocate the adoption of classroom-based research perspectives by practitioners. This need not mean carrying out full-scale research studies, which will often not be feasible. What it does mean is essentially to pursue an interest in finding out as much as possible about what students are learning, how their understanding and command of language is developing, what problems seem to persist, and so on, as systematically as working circumstances allow. Once again, a considerable challenge arises both for educational managers and for teachers themselves, as existing constraints ranging from heavy timetables to established role relations and mindsets can offer many obstacles to reflective teaching by individuals and groups. Any full evaluation of an educational programme must include the scope and direction that it is able to provide for teacher development through exploratory investigations of classroom learning experiences.

While formal research methodology is not a direct concern of this course, some understanding of basic research concepts is essential both to follow the professional literature and to pursue one's own classroom interests and activities in informed ways. Two fundamental concepts that are normally presented in introductory texts on language testing are those of reliability (can we trust the measurements to be accurate?) and validity (have we measured the abilities that we set out to measure?). A very brief introduction to these concepts now follows.

Although often associated narrowly with testing, the concepts of reliability and validity are also highly relevant to programme evaluations, and indeed to all empirical research procedures. As Nunan (1992) succinctly observes:

Reliability refers to the consistency of the results obtained from a piece of research. *Validity*, on the other hand, has to do with the extent to which a piece of research actually investigates what the researcher purports to investigate.

Nunan (1992:14)

Reliability is a necessary element of validity, but not a sufficient one. If our measurements are highly unreliable, we cannot draw any conclusions from them. On the other hand, even if our measurements are reliable, we also have to be confident that we have measured what actually interests us (e.g. reading comprehension ability, or level of student interest in a learning task) and not something else (e.g. writing ability, or the extent

to which students wish to please the teacher). While this distinction is clear in principle, it can become tricky in practice (Bachman, 1990: 238-241). For example, a valid test of reading comprehension would need to measure this ability, yet it can be difficult in practice to be clear about what constitutes the ability or "trait" being tested and what are features of a particular test method (e.g. the use of short-answer questions).

Validity is best considered as a unitary construct, though with many facets, according to Bachman (1990), who cites in support the established views of the American Psychological Association. Validity can nonetheless be estimated in a number of ways, which are sometimes listed by other writers as different kinds of validity. In such cases, the term "construct validity" is often used as a superordinate expression, as the overall aim is to justify a test in terms of the construct or model of the abilities that it seeks to measure. "Content", "concurrent" and "predictive" forms (or facets) of validity are often discussed in language testing. The first of these is concerned with whether a test adequately covers a body of knowledge or range of abilities (content), while the last two are both cases of criterion-relatedness, meaning that one measure is being validated through being compared to another which is either concurrently or subsequently administered.

"Face validity", a term indicating what a measure looks like, especially to non-specialists, can be discussed with a good deal of vigour. Clearly the fact that a test appears valid to some people offers no guarantee that it actually is so. To this extent, scepticism about the notion of face validity is wholly reasonable. Some testing specialists would reject the very notion of face validity for this reason. On the other hand, a test that does *not* appear to non-specialists to be a valid measure of what it claims to test can lead to problematic reactions on the part of test takers or other parties concerned with the results. It follows that, whatever their terminological preferences, language testers cannot afford to ignore or dismiss test appearance as an aspect of test validation.

If we attempt to relate notions of validity to language programme evaluation, we again see the importance of content and of criterion-relatedness. In terms of content, how far does an evaluation do justice to the full range of activities and experiences comprising a language programme? In terms of criterion-relatedness, how does one set of findings, such as questionnaire data, compare with other measures, such as comments made in learner diaries or during interviews? Although it is

not possible in a few words to do even summary justice to the complexities involved, it is likely that different measures of what is ostensibly the same phenomenon (such as student attitudes towards a learning task) will often yield rather different results. (Similar problems for validity arise in language testing when attempts are made to measure the same trait, such as "reading comprehension", by different test methods.) As for face validity, the luxury of discounting appearance to non-specialists is clearly both unacceptable and self-defeating in evaluation studies, since these have to be oriented towards decisions in which non-specialists often have a considerable say. Nevertheless, an appearance of validity alone is clearly not a sufficient basis for properly informed recommendations and decisions to be made.

The literature on language programme evaluation sometimes appears less concerned with statistical reliability, other than in the case of test results. However, the notion of reliability can usefully be extended to such areas as the percentage of inter-rater agreement on a content analysis of interview transcripts or of open-ended questionnaire data.

Before concluding this chapter, a brief illustration is offered of the relationship between various approaches to test validity. As the testing of oral abilities is often complex and time-consuming, attempts are sometimes made to produce highly indirect tests, such as paper-and-pencil responses to "what would you say in the following situations?", to estimate oral proficiency. Suppose that such attempts extend to an entire examination, as opposed to just one class activity among many: it then becomes essential to validate such an approach to testing. How, though, is this to be done? Some kinds of validation studies, besides establishing the internal consistency (reliability) of the paper-and-pencil tests in question, would seek to compare performance on these tests with performance during oral interviews (a criterion-related approach to validity). Even if the outcome of such comparisons were judged satisfactory, however, fundamental questions would still remain about the content and construct validity of paper-and-pencil tests as measures of oral abilities. Such tests are more likely to measure other abilities, such as aspects of knowledge about language varieties, than to measure oral abilities themselves, even if the two sets of abilities happen to be highly correlated. For such reasons, test validation becomes a matter of informed judgement that must draw on multiple perspectives, and not merely a question of establishing "the facts".

Where arguments about construct validity are motivated by explicit descriptions of abilities, such as oral communicative abilities, that a test fails to embody, they are clearly acceptable in test validation. In contrast, without such explicitness, otherwise comparable reactions on the part of non-specialists can easily be dismissed by specialists as mere appeals to face validity. This incurs a danger that potentially important information will be overlooked. While calls for explicitness deserve respect, people concerned with classroom language testing in particular cannot afford to ignore the impressions, however inarticulate (or opinionated) they may appear, of the people most concerned with the tests being set and taken. Indeed, eliciting impressions as explicitly and reflectively as possible from learners and others is a vitally important aspect of programme evaluation in general, and of classroom test validation as one aspect of this.

It is worth emphasising for this course that classroom language tests form part of a curriculum, and can themselves be studied from the perspectives of language programme evaluation. (See also Alderson *et al.*, 1995, which focuses largely on evaluating examinations.) Links between testing and teaching are of obvious importance in classroom contexts. In making use of tests, language teachers will be concerned about messages given to learners and other non-specialists, and the importance or otherwise that learners will subsequently attach to different kinds of activities in class. The crucial notion of washback effect (called backwash effect by some writers) is a name given to the impact of testing practices on teaching and learning. Returning to our earlier example, the washback effect of a complete paper-and-pencil examination of oral abilities would typically be highly detrimental to the direct encouragement of classroom oral interaction as this would not "count" in the eyes of most learners. This rather bold claim, naturally, must remain subject to classroom experiences and findings in particular situations.

Washback effect is of central importance to the relationship between classroom language testing (chapter 5) and programme evaluation (chapter 4). Washback effect is widely acknowledged in language testing handbooks as a crucial concern, but has often received short shrift in other discussions of test validation, and has not been a focus for serious empirical investigation until quite recently. One reason may be that language testing discussions of test validation have tended in the past to be more concerned with the relationship between particular tests and a

principled characterisation of language ability (chapter 3) than with the educational impact of testing practices.

Specialist concerns over what account should be given of language ability, and wider concerns over testing practices and their educational consequences, are both important for our purposes in this course. Explicit accounts of language ability and its development, in relation to contexts of learning and use, will evidently be helpful if we hope to evaluate the effectiveness of programmes that set out to enhance this ability. As a separate matter, the views of test takers and test users constitute essential data for the validation of forms of testing as aspects of educational practice. Test taker and user views are just one instance, though certainly an important one, of the attitudes and perceptions that classroom language learners and others bring to a language programme as a whole. There are both ideological and educational reasons to view learners as participants in language curricula, rather than as recipients of knowledge, so the views and values of participants are important in their own right as well as for their effects on learning practices. The general importance for our purposes of "lay" as well as "specialist" views about language abilities and language standards, in and beyond the language classroom, is the principal theme of chapter 2.

Activity 1.4 - Consulting learners

- In a teaching-learning situation known to you, in what ways do teachers form impressions of what learners understand, enjoy, think and do in the learning situation?
- What possible ways are not used? Why not? Would they be inappropriate for some reason, or are they too time-consuming, or do teachers just not think about them? Note as many suggested ways of consulting learners as you can.

1.5 Chapter summary and course overview

This introductory chapter is part of the overview taken in Part One (Contexts and Perspectives). It states the main aim of the course, which is to encourage all concerned with language education to explore and develop the educational relationships, both actual and potential, between testing, evaluation and language teaching. While introducing some key terms, the chapter emphasises the value-laden nature of concepts and issues in this field of discourse. The discussion encourages the adoption of classroom based research perspectives, not least as these can help to give focus and relevance to both investigative and speculative enquiry. The chapter also insists on the need for research and discussion to be open to diverse viewpoints and values, and to situate classroom practices in their wider educational and social perspectives.

A brief overview of the remaining chapters now follows.

Chapter 2 takes a broad look at issues in evaluating language, and picks up some implications for the goals and practices of a language curriculum.

Chapter 3 focuses on ways of describing and measuring language ability, including the development of applied linguistic models for these purposes, and again raises implications for language curricula.

Chapters 4 and 5 introduce some basic perspectives in evaluating language curricula and in testing in language classes.

In Part Two (Issues and Practices), there is a greater emphasis on moving from exposition through activities into practical explorations of issues that could be developed into small-scale research projects.

Chapters 6 and 7 outline some of the main ways in which language programme evaluations and language tests can be designed.

Chapters 8, 9 and 10 focus more closely on three areas: skills and subskills, tasks and genres, and performance criteria and rater judgements. These are chosen for their curricular relevance, their importance in both testing and programme evaluation, and their potential for interesting studies in specific teaching situations.

Chapter 11 reviews the course and relates its concerns to notions of professionalism in language teaching.

Further reading (Chapter 1)

Language testing has long been well served for a range of introductory texts, some relatively recent instances being Alderson, Clapham and Wall (1995), Bachman (1990), Bachman and Palmer (1996), Henning (1987), Hughes (1989) and Weir (1990; 1993). These books devote varying amounts of space to the role that testing can play in the evaluation of language programmes, but generally treat evaluation (rightly so, from their perspective) as just one of the purposes that language testing can serve.

Until recently, there have been very few texts on the evaluation of language teaching programmes, as Alderson and Beretta (1992) point out, though there is a substantial literature on educational evaluation in general. Among recent texts, Weir and Roberts (1994) is more oriented towards evaluations as professional activities undertaken by external appointees, while Lynch (1996) treats philosophical and procedural issues in "language program evaluation" as a focus of applied linguistic enquiry. In what is probably the most accessible introduction for teachers, Rea-Dickens and Germaine (1992) have tended to minimise the role of language testing and to maximise other aspects of educational evaluation in classroom and school. This is a natural reaction to the potentially reductive perspectives taken on evaluation from a language tester's starting point, but its effect is still to perpetuate a separate treatment of testing and evaluation as concerns in language teaching. (Brown, 1994, offers a fuller review of Rea-Dickens and Germaine's text.) The present book, in contrast, sets out to examine actual and possible relationships between the two areas, and so to assess the viability of the combined area of "language testing and evaluation" that others have sought to propose.

The relationship between evaluation and decision-making is quite widely discussed, with some variation in emphasis and scope. For example, Nunan (1992: 184-189) links evaluation to decision-making, to value judgements and to action. Nunan argues that an approach to evaluation that is limited to assessment data (let us look at students' test results to see how well course objectives have been achieved) is a narrow product-oriented view that takes no account of other relevant issues. We would also want to know whether test results reflect learning on the course itself or elsewhere, and to ask what other outcomes the course may have had; nor is it obvious that all course objectives are or should be actually reducible to behavioural outcomes.

Murphy (1985) maintains that essentially "evaluation is about making judgements, assessing the value and quality of what is done" (p. 2), but does not extend this directly to decision-making. Evaluation, in his view, may serve as input to decision-making, but these two processes are not the same thing. This is compatible with Bachman's claim that evaluation is involved when information is used as a basis for making decisions, but it stops short of Nunan's picture of evaluation as itself a decision-making process. Who is right, and does the point matter? In practice, the relationship between an evaluation procedure and decisions that follow will depend on who is carrying out an evaluation, and on whose behalf this is done. Where we evaluate our own teaching, for example, there seems little reason to distinguish sharply between our evaluations of present practice and our decisions about how best to proceed next. External evaluators, on the other hand, are not in the business of making decisions for their clients.

Henning (1987) lists "program evaluation" as one common use of tests, but does not consider it more widely; nor is language teaching a prominent theme in his book. In contrast, Rea-Dickens and Germaine (1992), while addressing language teaching concerns, devote the bulk of their book on evaluation to measures and procedures other than formal testing. It can be argued that Rea-Dickens and Germaine actually underrate the potential contribution of tests to evaluation procedures (see comments by Brown, 1994). Without pursuing this issue, we must again note that our combined theme of language testing and evaluation involves a linkage between two fields of activity, each with its own traditions in education.

A technical observation about test validation is that only certain approaches, notably the criterion ones, will yield a statistical measure of validity; others such as content analysis are not generally empirical in this way (Henning, 1987: 94ff). Test reliability, on the other hand, is normally computed as a statistic. Henning's text gives a good though somewhat technical account of different forms of reliability computation: test-retest (learners take the same test twice, and results are compared); parallel forms (results on two supposedly equivalent forms of a test are compared); inter-rater reliability (ratings by different assessors are correlated); split half reliability (a test is split into two supposedly comparable halves and performance on each half is compared; this is a

crude measure of internal consistency of a test); Kuder-Richardson formula 20 and 21 (more sophisticated measures of internal consistency).

The select annotated bibliography summarises some of the most useful sources on language testing and evaluation, and the full bibliography should help to suggest further reading depending on readers' own interests and backgrounds. Recent trends in language testing are also reviewed in several contributions in Alderson and North (1991) and Anivan (1991a), and by Weir (1990, 1993). Weir (1990) offers a thorough and accessible introduction to reliability and validity in language testing. The treatment in Bachman (1990) is rigorous but well worth following, and is taken up again in Bachman and Palmer (1996). Underhill (1987), in looking at oral testing from a teaching perspective, offers a refreshingly critical view of the "patronising" tendency among professional language testers to dismiss "face validity" as being spurious and of no account to them. In the case of language programme evaluation, a key requirement is that an evaluation report should actually help to guide decisions and action: Alderson (1992) develops this theme. It follows that appearance and presentation are crucial elements, though obviously not sufficient in themselves, when we assess the validity of programme evaluation studies. On programme evaluation as it relates to other aspects of curriculum design, see also Hutchinson and Waters (1987).