1 Origins and overview

This book is written for teachers of any language who are responsible for drawing up tests of language ability and for other professionals who may not be actively involved in teaching but who have some need to construct or evaluate language tests or examinations, or to use the information that such tests provide. (Since the distinction between a *test* and an *examination* is so vague, we use the terms interchangeably in this book.) Although our examples are mostly taken from the field of English as a Foreign Language, the principles and practice we describe apply to the testing of any language, and this book is certainly relevant to teachers and testers of any second or foreign language as well as to teachers and testers of first languages.

Those who are teaching may have to design placement tests for new incoming students, they may need to construct end-of-term or mid-year achievement tests for different levels within an institution, or they may be responsible for the production of major achievement test batteries at the end of a relatively long period of study.

Those who are not teaching but need to know how to produce tests include officials working for examination boards or authorities, and educational evaluators, who need valid and reliable measures of achievement.

Others who may need to design language tests include postgraduate students, researchers and academic applied linguists, all of whom need tests as part of their research. The test may be a means of eliciting linguistic data which is the object of their study, or it may be intended to provide information on linguistic proficiency for purposes of comparison with some other linguistic variable.

But in addition to those who need to construct tests, there are those who wish to understand how tests are and should be constructed, in order better to understand the assessment process, or in order to select from among a range of available tests one instrument suitable for their own contexts. Such people are often uncertain how to evaluate the claims that different examining authorities make for their own instruments. By understanding what constitutes good testing practice and becoming aware of current practices, such readers should be enabled to make more informed choices to suit their purposes.

In this book, we describe the process of test construction, from the

drafting of the initial test specifications through to the reporting of test scores and the devising of new tests in the light of developments and feedback. The book is intended to describe and illustrate best practice in test development, and the principles of test design, construction and administration that underpin such best practice.

The book is divided into eleven chapters, each dealing with one stage of the test construction process. Chapter 2 deals with the drawing up of the specifications on which the test is based. Chapter 3 describes the process of writing individual test items, their assembly into test papers and the moderation or editing which all tests should undergo. Chapter 4 discusses the importance of trialling the draft test and describes how tests should be analysed at this stage. Chapter 5 describes the training of markers and test administrators, whilst Chapter 6 shows how to monitor examiner reliability. Chapter 7 deals with issues associated with the setting of standards of performance and the reporting of results, whilst Chapter 8 describes further aspects of the process of test validation. Chapter 9 describes how reports on the performance of the test as a whole should be written and presented, and Chapter 10 discusses how tests can be developed and improved in the light of feedback and further research. The final chapter discusses the issue of standards in language testing and describes the current State of the Art.

Doubtless, this brief sketch of the content of the book sounds daunting: the test construction process is fairly complex and demanding. However, we have attempted to render our account user-friendly by various means. Each chapter opens with a brief statement of the questions that will be addressed and concludes with a checklist of the main issues that have been dealt with which can be consulted by busy teachers, exam board officials, researchers and test evaluators.

Our descriptions of the principles and procedures involved in language testing do not presuppose any knowledge of testing or of statistics. Indeed, we aim to provide readers with the minimum technical knowledge they will need to construct and analyse their own tests or to evaluate those of others. However, this is not a textbook on psychometrics: many good textbooks already exist, and the reader who becomes interested in this aspect of language testing is encouraged to consult the volumes listed at the end of this chapter. However, the reader should note that many books on educational measurement do not confine themselves to language testing, and they frequently assume a degree of numeracy or a familiarity with statistical concepts that our experience tells us most people involved in language testing do not possess. Our hope, though, is that having read this volume, such people will indeed be ready to read further.

Something we do not do in this book is to describe language testing

techniques in detail. This is partly because this topic is already addressed to some extent by a number of single volumes, for example, Oller 1979; Heaton 1988; Hughes 1990; Weir 1990; Cohen 1994. However, more importantly for us, we believe that it is not possible to do justice to this topic within the covers of one volume. In order to select test techniques and design good test items, a language tester needs a knowledge of applied linguistics, language teaching and language learning which cannot adequately be conveyed in a 'How-To' book, much less in the same volume as the discussion of testing principles and procedures. For the present we refer readers to the above language testing textbooks if what they need is a brief exemplification of test techniques.

Throughout the book we complement our discussion of the principles of test design with examples of how EFL examination boards in the United Kingdom implement these in practice. The second half of each chapter provides an illustration of how what we describe in the first part of each chapter is actually put into practice by examination boards in the UK.

Our aim is not to advocate that all tests should be constructed in the way UK examination boards do so: far from it. Rather, we wish to provide concrete examples that should help our readers understand the theory. We intend that this illustration should be relevant to all our readers and not just to exam board officials, although we believe that such officials will find it instructive to see the procedures and practices of other examination boards. Although the examples in this book are clearly located in a particular context – the UK – we know from experience that similar practices are followed elsewhere, and we firmly believe that language testers anywhere in the world will find aspects of the practice in a particular setting of relevance to their own context. The principles are universal, even if the practice varies.

We have discovered, from conducting workshops around the world with budding language testers, that anyone interested in learning about test construction, be it a placement test, an achievement test or a proficiency test, can learn from the experience of others. We present the data on current practice in the UK critically: we discuss strengths and weaknesses, and make suggestions for change if best practice is to be realised. The reader can perhaps take heart that even examination boards do not always do things perfectly; we all have things to learn from relating principles to practice.

We gathered this information in a variety of ways which we describe below, but first we digress to describe why this volume came to be written. All three authors had experienced considerable frustration at not having available any account of how examination boards construct

language tests. We have all three taught language testing on MA courses, in-service courses for practising teachers, and in workshops around the world for different audiences. We have had considerable experience of working with UK examination boards as item writers, members of editing committees, examiners, test validators and testing researchers. We are all acquainted with language testing theory and the principles of test design. Yet nowhere had we found an adequate description of how examinations are constructed in order to implement the principles.

Our first attempt systematically to collect information about UK examination boards began in 1986, when we were invited to carry out a research project that was to make recommendations for quality control procedures in new English language examinations in Sri Lanka. We held a series of interviews with representatives of various EFL examining boards in order to find out how they conducted tests of writing and speaking. These interviews resulted in a number of reports whose content was subsequently agreed with respondents. The reports were circulated internally at Lancaster and made available to visitors and students, but were never published, and did not in any case cover all the bodies engaged in EFL examining in the UK.

One of the authors of this book was invited by Karl Krahnke and Charles Stansfield to contribute as co-editor to the TESOL publication *Reviews of English Language Proficiency Tests*. Part of the work involved commissioning reviews of twelve UK EFL examinations. These reviews were subsequently sent to the respective examination boards for comment. They were then amended where necessary and published in Alderson et al. 1987. Many of the reviewers made similar points about both the strengths and weaknesses of UK exams, some of which were contested by the examination boards. Of the twelve UK tests reviewed, the reviewers criticised nine for failing to provide sufficient evidence of reliability and validity, and in only two cases did the reviewers express satisfaction with the data provided. Alderson included in this TESOL publication *An Overview of ESL/EFL Testing in Britain*, which explained British traditions to readers from other countries. In this overview he stated:

> Due to the constant need to produce new examinations and the lack of emphasis by exam boards on the need for empirical rather than judgemental validation, these examinations are rarely, if ever, tried out on pupils or subjected to the statistical analyses of typical test production procedures. Examination boards do not see the need to pretest and validate their instruments, nor conduct post-hoc analyses of their tests' performance. Although the objective items in the tests are

> usually pretested, the statistics are rarely published.
>
> (Alderson et al. 1987)

This overview was subsequently updated for a chapter in Douglas 1990 on UK EFL examining. In order to gather up-to-date information, Alderson sent a copy of the original overview to UK examination boards and asked whether it was still substantially correct or whether any amendments were necessary. Few boards responded, but those that did said that things had not changed.

The Lancaster Language Testing Research Group next decided to survey the boards. For this purpose, we referred to the Appendix in Carroll and West 1989, the report of the English Speaking Union (ESU)'s Framework Project. In addition, we decided to include in our survey the Schools Examination and Assessment Council (SEAC, formerly SEC, the Secondary Examinations Council), a body set up by the Government and charged with the responsibility of establishing criteria for judging educational examinations and of determining the validity of such exams.

Our survey was in three parts. First, in December 1989 we wrote letters to each of the examining authorities listed, and to SEAC. These letters contained the three following open-ended questions, which sought to elicit the boards' own views of their standards and the procedures they used to establish reliability and validity:

1. Do you have a set of standards to which you adhere?
2. What procedures do you follow for estimating test reliability?
3. What procedures do you follow to ensure test validity?

We presented the results of this first phase of our research to a meeting of the Association of British ESOL Examining Boards (ABEEB) in November 1990.

Secondly, we circulated a questionnaire to the same examination boards in December 1990. A summary of the responses to this questionnaire forms part of the second half of each chapter of this book. A written version of the results was circulated to the responding examination boards for their comments in May 1991, and discussions were held about the study. We subsequently gave each board the opportunity to update its response, on the grounds that much development had taken place during the intervening months, and we received very detailed responses to this from the University of Cambridge Local Examinations Syndicate (UCLES) in particular.

Thirdly, we also received a large amount of printed material associated with the various examinations from the boards, and we have analysed these in some detail: we present summaries of and examples from this analysis where appropriate in each chapter. It may be of

interest to the reader, however, to know which documents we received. They are listed, together with the names of the boards and the examinations they produce, in Appendix 1.

A summary of some of the main results from Phase Two of the survey has already appeared in Alderson and Buck 1993, but this book contains more detail than that paper, and updates much of the information in it. It is, of course, possible there may have been changes in the procedures followed by some boards since we completed our research. We hope that we have not misrepresented any examining body, but would welcome any corrections, additions or other modifications that might be necessary. Since most examination boards preferred to remain anonymous when the results of the survey were published, we only name those boards which gave us permission to do so, or where we are quoting from publicly available literature.

This book has very much benefited from the knowledge gained as a result of the survey. We hope that our readers will benefit equally from being able to read an account of current practice alongside a description of the principles of language testing, and the procedures we believe to be appropriate for test construction.

More important than the details of the practice of individual examination boards are the principles that should underlie language testing practice, and that is why each chapter contains a detailed treatment of these principles. That is also why each chapter ends with a section that lists the questions an evaluator might ask of any test, or a checklist of things that test designers or evaluators need to pay attention to.

The overarching principles that should govern test design are *validity* and *reliability*, and we make constant reference to these throughout the book. Validity is the extent to which a test measures what it is intended to measure: it relates to the uses made of test scores and the ways in which test scores are interpreted, and is therefore always relative to test purpose. Although the only chapter in the book with a reference to validity in its title is Chapter 8, the concept of validity is central to all the chapters. Reliability is the extent to which test scores are consistent: if candidates took the test again tomorrow after taking it today, would they get the same result (assuming no change in their ability)? Reliability is a property of the test as a measuring instrument, but is also relative to the candidates taking the test: a test may be reliable with one population, but not with another. Again, although reliability is only mentioned in one chapter title (Chapter 6), it is a concept which runs through the book.

We attempt to define the specialist testing terminology when we first use it, and so we will not enter into further definitions at this point.

However, we have supplied a glossary of important terms in testing, for the reader's reference. We are also aware that most readers will not be familiar with the abbreviations and acronyms often used in EFL testing, and in particular those that are used to denote the UK examination boards. We have therefore also supplied a comprehensive list of such terms at the end of the book.

The research reported in this book is the result of many months of collaboration amongst members of the Lancaster Language Testing Research Group and visiting researchers. We are very grateful to the following for their assistance, encouragement and criticisms: Joan Allwright, Gary Buck, Nicki McLeod, Frank Bonkowski, Rosalie Banko, Marian Tyacke, Matilde Scaramucci and Pal Heltai. We would also like to thank the various examination boards, the British Council, and Educational Testing Service, New Jersey, for their help.

Bibliography

Alderson, J.C. and G. Buck. 1993. Standards in Testing: A Survey of the Practice of UK Examination Boards in EFL Testing. *Language Testing* 10(1): 1–26.

Alderson, J.C., K. Krahnke and C. Stansfield (eds.). 1987. *Reviews of English Language Proficiency Tests*. Washington, DC: TESOL.

Anastasi, A. 1988. *Psychological Testing*. London: Macmillan.

Carroll, B.J. and R. West. 1989. *ESU Framework: Performance Scales for English Language Examinations*. London: Longman.

Cohen, A. 1994. *Assessing Language Ability in the Classroom*. 2nd edition. Rowley, Mass.: Newbury House/Heinle and Heinle.

Crocker, L. and J. Algina. 1986. *Introduction to Classical and Modern Test Theory*. Chicago, Ill.: Holt Rinehart Winston.

Douglas, D. (ed.). 1990. *English Language Testing in U.S. Colleges and Universities*. Washington, DC: NAFSA.

Ebel, R.L. 1979. *Essentials of Educational Measurement*. 3rd edition. Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R.L. and D.A. Frisbie. 1991. *Essentials of Educational Measurement*. 5th edition. Englewood Cliffs, NJ: Prentice-Hall.

Guilford, J.P. and B. Fruchter. 1978. *Fundamental Statistics in Psychology and Education*. Tokyo: McGraw Hill.

Hambleton, R.K., H. Swaminathan and H.J. Rogers. 1991. *Fundamentals of Item Response Theory*. Newbury Park, Calif.: Sage Publications.

Heaton, J.B. 1988. *Writing English Language Tests*. 2nd edition. London: Longman.

Henning, G. 1987. *A Guide to Language Testing*. Cambridge, Mass.: Newbury House.

Hughes, A. 1990. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Ingram, E. 1977. Basic Concepts in Testing. In J.P.B. Allen and A. Davies (eds.), *Testing and Experimental Methods*. Oxford: Oxford University Press.

Lord, F.M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Oller, J.W. 1979. *Language Tests at School*. London: Longman.

Popham, W.J. 1990. *Modern Educational Measurement: A Practitioner's Perspective*. 2nd edition. Boston, Mass.: Allyn and Bacon.

Weir, C.J. 1990. *Communicative Language Testing*. Englewood Cliffs, NJ: Prentice-Hall Regent.

2 Test specifications

The questions that this chapter seeks to answer in detail are: What are test specifications? Who needs test specifications? What should test specifications look like? How can we draw up test specifications? What do current EFL examinations prepare in the way of specifications?

2.1 What are test specifications?

A test's specifications provide the official statement about what the test tests and how it tests it. The specifications are the blueprint to be followed by test and item writers, and they are also essential in the establishment of the test's construct validity.

Deriving from a test's specifications is the test syllabus. Although some UK examination boards use *specifications* and *syllabus* interchangeably, we see a difference between them. A test specification is a detailed document, and is often for internal purposes only. It is sometimes confidential to the examining body. The syllabus is a public document, often much simplified, which indicates to test users what the test will contain. Whereas the test specification is for the test developers and those who need to evaluate whether a test has met its aim, the syllabus is directed more to teachers and students who wish to prepare for the test, to people who need to make decisions on the basis of test scores, and to publishers who wish to produce materials related to the test.

The development and publication of test specifications and syllabuses is, therefore, a central and crucial part of the test construction and evaluation process. This chapter will describe the sorts of things that test specifications and syllabuses ought to contain, and will consider the documents that are currently available for UK EFL tests.

2.2 Who needs test specifications?

As has already been suggested, test specifications are needed by a range

of different people. First and foremost, they are needed by those who produce the test itself. Test constructors need to have clear statements about who the test is aimed at, what its purpose is, what content is to be covered, what methods are to be used, how many papers or sections there are, how long the test takes, and so on. In addition, the specifications will need to be available to those responsible for editing and moderating the work of individual item writers or teams. Such editors may operate in a committee or they may be individual chief examiners or board officials. (See Chapter 3 for further discussion of the editing process.) In smaller institutions, they may simply be fellow teachers who have a responsibility for vetting a test before it is used. The specifications should be consulted when items and tests are reviewed, and therefore need to be clearly written so that they can be referred to easily during debate. For test developers, the specifications document will need to be as detailed as possible, and may even be of a confidential nature, especially if the test is a 'high-stakes' test.

Test specifications are also needed by those responsible for or interested in establishing the test's validity (that is, whether the test tests what it is supposed to test). These people may not be the test constructors, but outsiders or other independent individuals whose needs may be somewhat different from those of the item writers or editors. It may be less important for validators to have 'practical' information, for example, about the length of the test and its sections, and more important to know the theoretical justification for the content: what theories of language and proficiency underpin the test, and *why* the test is the way it is.

Test users also need descriptions of a test's content, and different sorts of users may need somewhat different descriptions. For example, teachers who will be responsible for the learners placed in their classes by a test need to know what the test scores mean: what the particular learners know, what they can do, what they need to learn. Although the interpretation of test scores is partly a function of how scores are calculated and reported (see Chapter 7), an understanding of what scores mean clearly also relates to what the test is testing, and therefore to some form of the specifications.

Teachers who wish to enter their students for some public examination need to know which test will be most appropriate for their learners in relation to the course of instruction that they have been following. They need information which will help them to decide which test to choose from the many available. Again, some form of the specifications will help here – probably the simplified version known as the syllabus.

Admissions officers who have to make a decision on the basis of test