

Testing for Language Teachers

Third Edition

Arthur Hughes
Jake Hughes



CAMBRIDGE

Testing for Language Teachers

Third Edition

Arthur Hughes and Jake Hughes

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781108714822

© Cambridge University Press 2020

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2020

20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Printed in Great Britain by Ashford Colour Press Ltd.

A catalogue record for this publication is available from the British Library

ISBN 978-1-108-71482-2 Paperback

ISBN 978-1-108-71485-3 Apple iBook

ISBN 978-1-108-71483-9 ebooks.com eBook

ISBN 978-1-108-71487-7 Google eBook

ISBN 978-1-108-71486-0 Kindle eBook

Cambridge University Press has no responsibility for the persistence or
accuracy of URLs for external or third-party internet websites referred to in
this publication, and does not guarantee that any content on such websites is,
or will remain, accurate or appropriate.

Dedication

For Vicky, Meg, Nerys, Arlo and Elias

Contents

Dedication	iii
Acknowledgements	vii
Preface	ix
1 Testing, teaching and society: the language tester's responsibilities	1
2 Testing as problem solving: an overview of the book	8
3 Kinds of tests and testing	11
4 Validity	29
5 Reliability	40
6 Achieving positive backwash	57
7 Stages of test development	63
8 Common test techniques	78
9 Testing writing	87
10 Testing speaking	115

11	Testing reading	140
12	Testing listening	163
13	Testing grammar and vocabulary	176
14	Testing overall ability	192
15	Tests for young learners	206
16	Beyond testing: other means of assessment	227
17	New technology and language testing	233
18	Test administration	238
19	The statistical analysis of test data	242
	Appendix 1 Item banking	256
	Appendix 2 Checklist for teachers choosing tests for their students	258
	Appendix 3 The secrets of happiness	260
	Bibliography	261
	Author index	278
	Subject index	282

Acknowledgements

The authors and publishers acknowledge the following sources of copyright material and are grateful for the permissions granted. While every effort has been made, it has not always been possible to identify the sources of all the material used, or to trace all copyright holders. If any omissions are brought to our notice, we will be happy to include the appropriate acknowledgements on reprinting and in the next update to the digital edition, as applicable.

Key: C = Chapter Text

C3: Text taken from 'IELTS band score descriptors'. Copyright © UCLES. Reproduced with kind permission of Cambridge Assessment English. Text taken from 'IELTS Band 7 descriptors'. Copyright © UCLES. Reproduced with kind permission of Cambridge Assessment English. Text taken from 'IELTS score to Band score conversion table'. Copyright © UCLES. Reproduced with kind permission of Cambridge Assessment English. Text taken from 'Interagency Language Roundtable Language Skill Level Descriptions - Reading'. Copyright © Interagency Language Roundtable. Reproduced with kind permission. **C9:** Text taken from *Cambridge English Qualifications B2 First Handbook for Teachers*. Copyright © UCLES. Reproduced with kind permission of Cambridge Assessment English. Story taken from *Guide to Pearson Test of English Young Learners Breakthrough (Level 4)*. Copyright © 2012 Pearson Education Ltd. Reproduced with kind permission. Graph taken from 'Academic Writing – task 1'. Copyright © takeielts.org. Reproduced with kind permission. Text taken from 'TOEFL iBT® Test Independent Writing Rubrics' by Educational Testing Service. Copyright © 2014 Educational Testing Service. TOEFL iBT® Test Independent Writing Rubrics are reprinted by permission of Educational Testing Service, the copyright owner. All other information contained within this publication is provided by Cambridge University Press and no endorsement of any kind by Educational Testing Service should be inferred. Text taken from 'ACTFL Proficiency Guidelines 2012 – Writing'. Copyright © 2012 American Council on the Teaching of Foreign Languages (ACTFL). Reproduced with kind permission. Text republished with permission of McGraw Hill from *Testing English as a second language* by David P. Harris. Copyright © 1969 McGraw Hill. Reproduced with kind permission via Copyright Clearance Center. Text taken from *Testing ESL Composition: A Practical Approach* by Holly L. Jacobs, Stephen A Zinkgraf, Deanna R. Wormuth and Jane B. Hughey. Copyright © 1981 Newbury House Publishers. Reproduced with permission. **C10:** Text taken from *Cambridge English Qualifications B2 First Handbook for Teachers*. Copyright © UCLES 2016. Reproduced with permission of Cambridge Assessment English. Text taken from 'Interagency Language Roundtable Language Skill Level Descriptions – Speaking'. Copyright © Interagency Language Roundtable. Reproduced with kind permission. **C11:** Text taken from 'Look on the bright side, banish the blues and think yourself happy' by Harry Wallop, 05.07.2013, *The Telegraph*. Copyright © 2013 Telegraph Media Group Limited. Reproduced with permission. Text taken from 'IELTS academic writing sample'. Copyright ©

UCLES. Reproduced with kind permission of Cambridge Assessment English. Text taken from 'The secrets of happiness' by Mihaly Csikszentmihalyi. Copyright © Mihaly Csikszentmihalyi. Reproduced with kind permission. Text taken from 'Dig in! Archaeologists serve up ancient menus for modern tables' by James Tapper, 11.08.2019, *The Guardian*. Copyright © 2019 *The Guardian*. Reproduced with permission. **C12:** Text taken from 'IELTS listening sample test questions'. Copyright © UCLES. Reproduced with kind permission of Cambridge Assessment English. Text taken from 'IELTS sample test questions'. Copyright © UCLES. Reproduced with kind permission of Cambridge Assessment English. Text taken from *C21 English for the 21st Century Level 3 course book*. Copyright © Garnet Publishing Ltd. Reproduced with kind permission.

C13: Text taken from *Complete First* by Guy Brook-Hart. Copyright © 2014 Cambridge University Press. Text taken from *Cambridge English B2 First Handbook*. Copyright © UCLES. Reproduced with kind permission of Cambridge Assessment English. **C14:** Text taken from *The Cloze Technique and ESL Proficiency* by John W. Oiler Jr and Christine A. Conrad. Language Learning, 21: 183-194. doi: 10.1111/j.1467-1770.1971.tb00057.x. Copyright © John Wiley & Sons, Inc. Reproduced with kind permission. Text taken from *A survey of research on the C-Test1* by Christine Klein-Braley and Ulrich Raatz, *Language Testing*, 1(2), 134 -146. <https://doi.org/10.1177/026553228400100202>. Copyright © 1984 SAGE Publications. Reproduced with kind permission. Text taken from 'C2 Proficiency (CPE) Practice Test' by Johanna Kieniewicz. Reproduced with kind permission. Text taken from 'UAB Cloze test'. Copyright © UAB Language Service. Reproduced with kind permission. **C15:** Table taken from *Assessing the Language of Young Learners* by Angela Hasselgreen and Gwendydd Caudwell. Copyright © 2016 Equinox Publishing Limited. Reproduced with permission.

Photography

The following photos are sourced from Getty Images.

C10: Alexa Miller/The Image Bank/Getty Images Plus; Morsa Images/DigitalVision; **C11:** IMNATURE/iStock/Getty Images Plus; Ogphoto/E+; curtoicurto/iStock/Getty Images Plus.

The following photos are sourced from other libraries/sources.

C10: Pres Panayotov/Shutterstock; Copyright © Simon Roberts; **C11:** Dr Keith Wheeler/Science Photo Library; plearn/Shutterstock; Krzysztof Kostrubiec/Shutterstock.

Cover photography by Purestock/Getty Images.

Illustrations

Illustrations taken from *Cambridge English Young Learners English Test Movers level sample papers*. Copyright © 2014 UCLES. Reproduced with kind permission of Cambridge Assessment English. Illustrations taken

from *Assessing the Language of Young Learners* by Angela Hasselgreen and Gwendydd Caudwell. Copyright © 2016 Equinox Publishing Limited. Reproduced with permission. Illustrations taken from *Cambridge English Young Learners sample papers*. Copyright © UCLES. Reproduced with kind permission of Cambridge Assessment English. Illustrations taken from *Primary Colours 1* by Diana Hicks and Andrew Littlejohn. Copyright © 2002 Cambridge University Press; Q2A Media Services Pvt. Ltd.

URL

The publisher has used its best endeavors to ensure that the URLs for external websites referred to in this book are correct and active at the time of going to press. However, the publisher has no responsibility for the websites and can make no guarantee that a site will remain live or that the content is or will remain appropriate.

Preface

We live in a world of accelerating change, change to which language testing is not immune. Since the publication of the second edition of this book, technological developments have led to more and more language tests being delivered online, with even the writing and speaking components of some of them being scored without human intervention. While not all of these developments are yet directly applicable to teacher-made tests, we believe that teachers should be aware of them, partly because they may soon become applicable, but also so that teachers can advise their students on the choice of commercially available tests. To this end, we have added a chapter on new technology and, in an appendix, a checklist to help teachers choose tests.

Despite changes, the principles underlying language testing remain the same. We believe that there is still a place for a straightforward introductory guide to the field. The objective of this book is unchanged: to help language teachers write better tests. It takes the view that test construction is essentially a matter of problem solving, with every teaching situation setting a different testing problem. In order to arrive at the best solution for any particular problem – the most appropriate test or testing system – it is not enough to have at one's disposal a collection of test techniques from which to choose. It is also necessary to understand the principles of testing and how they can be applied in practice.

It is relatively straightforward to introduce and explain the desirable qualities of tests: validity, reliability, practicality, and positive backwash; this last referring to the favourable influence that testing can have on teaching and learning. It is much less easy to give realistic advice on how to achieve them in teacher-made tests. One is tempted either to ignore the issue or to present as a model the not always appropriate methods of large-scale testing organisations. In resisting this temptation, we have made recommendations that we hope teachers will find practical but which we have also tried to justify in terms of language testing theory.

Exemplification throughout the book is from the testing of English as a foreign language. This reflects both our own experience in language testing and the fact that English will be the one language known by all readers. We trust that it will not prove too difficult for teachers of other languages to find or construct parallel examples of their own.

Because the objective and general approach of the book remain those of the second edition, much of the text remains. However, we have made changes throughout. As well as identifying and outlining significant new developments, we have added a more extended discussion of language testers' responsibilities. There are new chapters on non-testing methods of assessment and, as noted above, on new technology. Most examples have been replaced by more recent ones.

The number of published articles and books in the field continues to increase. Many – perhaps most – of the articles have been of a theoretical or technical nature, not directly relevant to the concerns of language teachers. Even where their relevance is clear, in order to keep the text accessible to newcomers to the field, we have usually restricted references to the Further reading sections. These sections are intended to act as a guide for those readers who wish to go more deeply into the issues raised in the book, and also to provide an outline of the state of language testing today. They also contain recommendations of a number of recent books which, in an accessible fashion, treat areas of language testing (such as the testing of a particular skill) in greater depth than is possible in the present volume. Also included in the Further reading section are references to useful resources that are available online. One such resource is worth mentioning here, since it represents a doorway into so many others. It is The Language Testing Resources Website, and is highly recommended.

We must acknowledge the contributions of others: MA and research students at Reading University, too numerous to mention by name; friends and colleagues, Paul Fletcher, Michael Garman, Don Porter, John Slaght, Tony Woods, and the late Cyril Weir; Angela Hasselgreen, who again shared thoughts on the testing of young learners and (along with Hilde Olsen) facilitated the provision of Norwegian materials; Janna Fox, who provided background information to one of her articles; Nick Saville, who discussed the possible content of this new edition at the outset and who promptly answered all our queries thereafter, even when abroad on holiday; Karen Momber, who gave us encouragement and support throughout; Christine Coombe, who read and commented on every chapter as we wrote it; Alison Sharpe, who made many helpful suggestions for improving the text and sought new examples for inclusion; Jo Timerick, for her support, especially in those tricky final stages; finally our wives – one of whom drew the cartoon series on page 219, and who proofread the entire text – who remained patient while we absented ourselves from family life for longer periods than they perhaps thought justified.

One last thing. We are father and son. Some articles were written and tests constructed by one of us when the other was still in primary school. For simplicity, however, we have referred to ourselves throughout as 'we'.

1 Testing, teaching and society: the language tester's responsibilities

The language tester has responsibilities to everyone who holds a stake in a test. Stakeholders include test-takers, teachers, parents, administrators, professional bodies, and many others; in fact, anyone involved with the test in any way. The higher the stakes in a test, the greater are the tester's responsibilities.

By high-stakes tests we mean tests which may have a significant effect on the test-takers' lives. Tests on which success is a prerequisite for university study abroad or for advancement in one's career are examples of high-stakes tests. This is where responsibility is greatest.

At the other end of the scale are classroom tests which may be designed solely to provide a teacher with information about students' grasp of what has recently been taught. But even here tests should be constructed in a responsible way.

What are the language tester's responsibilities? In brief, they are to:

1. write tests which give accurate measures of the test-takers' ability;
2. endeavour to make the impact of tests as positive as possible.

We shall treat each of these responsibilities in turn.

Accuracy

Language tests too often fail to measure accurately whatever it is that they are intended to measure. Teachers know this. Students' true abilities are not always reflected in the test scores that they obtain. To a certain extent this is inevitable. Language abilities are not easy to measure; we cannot expect a level of accuracy comparable to those of measurements in the physical sciences. But we can expect greater accuracy than is frequently achieved.

Why are tests inaccurate? The causes of inaccuracy (and ways of minimising their effects) are identified and discussed in subsequent chapters, but a short answer is possible here. There are two main sources of inaccuracy. The first of these concerns test content and test techniques. Let us take as an example the testing of writing ability. If we want to know how well someone can write, there is absolutely no way we can get a really accurate measure of their ability by means of a multiple choice test. Perhaps surprisingly, in the past professional testers in large organisations expended great effort, and not a little money, in attempts to

do just that. Why? It was in order to avoid the difficulty and expense of scoring hundreds of thousands of compositions. Accuracy was sacrificed for reasons of economy and convenience. In our view, the testers involved were failing to meet their responsibilities. Happily, the practice of testing writing ability using multiple choice items has been largely abandoned. Nowadays, students' scripts are delivered electronically to markers, and procedures are in place to ensure standardisation of scoring. However, the desire of large testing organisations to find more economical solutions to their testing problems remains. The scoring of written work solely by computers, which we will discuss in the chapter on the testing of writing, is an example of this.

While few teachers would ever have wished to test writing ability using multiple choice items, the continued use of that technique in large-scale, professional testing (for purposes other than to measure writing ability) tends to lead to their inclusion in teacher-made tests. In our experience, teachers' multiple choice items are often of a very poor standard. Good multiple choice items are notoriously difficult to write. A great deal of time and effort has to go into their construction. Too many multiple choice tests are written where the necessary care and attention are not given. The result is a set of poor items that cannot possibly provide accurate measurements. One of the principal aims of this book is to discourage the use of inappropriate techniques and to show that teacher-made tests can be superior in certain respects to their professional counterparts.

The second source of inaccuracy is lack of reliability. This is a technical term that is explained in Chapter 5. For the moment it is enough to say that a test is reliable if it measures consistently. With a reliable test you can be confident that someone will get more or less the same score, whether they happen to take it on one particular day or on the next; whereas on an unreliable test the score is quite likely to be considerably different, depending on the day on which it is taken. Unreliability has two origins. The first is the interaction between the person taking the test and features of the test itself. Human beings are not machines and we therefore cannot expect them to perform in exactly the same way on two different occasions, whatever test they take. As a result, we expect some variation in the scores a person gets on a test, depending on when they happen to take it, what mood they are in, how much sleep they had the night before. However, what we can do is ensure that the tests themselves don't increase this variation by having unclear instructions, ambiguous questions, or items that result in guessing on the part of the test-takers. Unless we minimise these features, we cannot have confidence in the scores that people obtain on a test.

The second origin of unreliability is to be found in the scoring of a test. Scoring can be unreliable, in that equivalent test performances are accorded significantly different scores. For example, the same composition may be given very different scores by different markers (or even by

the same marker on different occasions). Fortunately, there are ways of minimising such differences in scoring. Most (but not all) large testing organisations, to their credit, take every precaution to make their tests, and the scoring of them, as reliable as possible, and are generally highly successful in this respect. Small-scale testing, on the other hand, tends to be less reliable than it should be. Another aim of this book, then, is to show how to achieve greater reliability in testing. Advice on this is to be found in Chapter 5.

Multiple measures

There is a growing recognition that, however valid and reliable a single test may be, by itself it cannot be depended on to give an accurate picture of every individual candidate's ability. For this reason, there has been a move towards looking at more than one measure when taking decisions which may have important implications for people's lives. These different measures may be taken at different times, and so provide evidence of the progress that the candidate has been making towards the required standard. Of course, the mere fact that there are multiple measures of ability does not guarantee that an assessment based on them will be accurate. Much will depend on the accuracy of the different measures themselves. There are also issues as to how the measures should be combined in coming to a decision as to a candidate's ability.

Impact

The term *impact*, as it is used in educational measurement, is not limited to the effects of assessment on learning and teaching but extends to the way in which assessment affects society as a whole, and has been discussed in the context of the ethics of language testing.

Backwash

The impact of testing on teaching and learning is known as *backwash* (sometimes referred to as *washback*), and can be harmful or positive. If a test is regarded as important, if the stakes are high, preparation for it can come to dominate all teaching and learning activities. And if the test content and testing techniques are at variance with the objectives of the course, there is likely to be harmful backwash. An instance of this would be where students are following an English course that is meant to train them in the language skills (including writing) necessary for university study in an English-speaking country, but where the language test that they have to take in order to be admitted to a university does not test those skills directly. If the skill of writing, for example, is tested only by multiple choice items, then there is great pressure to practise such items rather than practise the skill of writing itself. This is clearly undesirable.

We have just looked at a case of harmful backwash. However, backwash can also be positive. One of us was once involved in the development of an English language test for an English-medium university in a non-English-speaking country. The test was to be administered at the end of an intensive year of English study there and would be used to determine which students would be allowed to go on to their undergraduate courses (taught in English) and which students would have to leave the university. A test was devised which was based directly on an analysis of the English language needs of first-year undergraduate students, and which included tasks as similar as possible to those which they would have to perform as undergraduates (reading textbook materials, taking notes during lectures, and so on).

The introduction of this test, in place of one which had been entirely multiple choice, had an immediate effect on teaching: the syllabus was redesigned, new books were chosen, classes were conducted differently. The result of these changes was that by the end of their year's training, in circumstances made particularly difficult by greatly increased numbers and limited resources, the students reached a much higher standard in English than had ever been achieved in the university's history. This was a case of positive backwash. The test, in new versions of course, is still in place more than thirty years later.

Davies (1968:5) wrote that 'the good test is an obedient servant since it follows and apes the teaching'. We find it difficult to agree. The proper relationship between teaching and testing is surely that of partnership. It is true that there may be occasions when the teaching programme is potentially good and appropriate but the testing is not; we are then liable to suffer from harmful backwash. This would seem to be the situation that led Davies in 1968 to confine testing to the role of servant to the teaching. But equally there may be occasions when teaching is poor or inappropriate and when testing is able to exert a positive influence. We cannot expect testing only to follow teaching. Rather, we should demand of it that it is supportive of good teaching and, where necessary, exerts a corrective influence on bad teaching. If testing always had a positive backwash on teaching, it would have a much better reputation among teachers. These days, most members of the testing community would probably agree with what we are saying. However, we include it because we know that there are teaching institutions throughout the world where the view expressed by Davies still persists. Chapter 6 of this book is devoted to a discussion of how positive backwash can be achieved.

Impact beyond the classroom

Language tests have an impact outside the teaching and learning environment. They are used to make decisions about employment, citizenship, immigration and the granting of asylum. There are two common problems with the way that tests are used for these purposes.

First, the tests are often inappropriate. For example, a test designed to measure language ability for university study is routinely used to determine whether nurses have sufficient English to work on hospital wards in the United Kingdom. One can be sure that nurses whose English is perfectly adequate for their work are nevertheless rejected because of their scores on that test. Professional bodies are often resistant to change (and what they see as avoidable expense). Several years ago, we were consulted by one august British body as to the appropriateness of an academic English test then being used for the measurement of the English ability of applicants. We advised that a modified version of a test specifically designed for their profession in another English-speaking country would give more accurate results. We were encouraged to think that this advice would be followed, only to see, while writing this chapter, that the old test was still in place. The only change was that higher grades were required!

Second, users of test scores, such as government agencies, typically act without awareness of the necessarily imprecise nature of those scores. Life-changing decisions are too often made on the basis of a single test score, even though the candidate score or grade is so close to the one required that no one can be confident that he or she does not have the language ability deemed necessary. The recognition of this has led to the introduction of multiple measures assessment in some contexts.

What should we do?

This book is meant for language teachers. It would be unreasonable to assign to them all the responsibilities that we have identified in this chapter. Nevertheless, we believe that teachers can play a more important part in language testing than they might expect.

If they begin by gaining a good understanding of the principles of language testing and familiarise themselves with good practice in the field (frequently referred to as *language assessment literacy* – see Further reading), they should be able to write better tests themselves. This will also allow them to enlighten others who are involved with the testing process within educational institutions. We believe that the better all of the stakeholders in a test or testing system understand testing, the better the testing will be and, where relevant, the better it will be integrated with teaching. The stakeholders we have in mind include test-takers, teachers, test writers, school or college administrators, education authorities and examining bodies. The more they interact and cooperate on the basis of shared knowledge and understanding, the better and more appropriate should be the testing in which they all have a stake. Teachers are probably in the best position to understand the issues, and then to share their knowledge with others.

Teachers with a good grasp of assessment can have a significant influence beyond the immediate educational system in which they operate. We have

referred more than once to the testing of writing ability through multiple choice items. This was the practice followed by those responsible for *TOEFL*[®] (Test of English as a Foreign Language) – the test taken by most non-native speakers of English applying to North American universities. Over a period of many years they maintained that it was simply not possible to test the writing ability of hundreds of thousands of candidates by means of a composition: it was impracticable and the results, anyhow, would be unreliable. Yet in 1986 a writing test (Test of Written English), in which candidates actually have to write for thirty minutes, was introduced as a supplement to *TOEFL*[®]. The principal reason given for this change was pressure from English language teachers who had finally convinced those responsible for the *TOEFL*[®] of the overriding need for a writing task that would provide positive backwash.

We believe that the power of social media and the ease of creating online petitions will only strengthen teachers' influence on the nature and use of language tests in society.



READER ACTIVITIES

1. Think of tests with which you are familiar (the tests may be international or local, written by professionals or by teachers). What do you think the backwash effect of each of them is? Harmful or positive? What are your reasons for coming to these conclusions?
2. Consider these tests again. Do you think that they give accurate or inaccurate information? What are your reasons for coming to these conclusions?
3. Find the ILTA Code of Ethics and Guidelines online. Which elements in these seem most relevant to your testing situation (or one you are familiar with)? Do you see any problems in their application?
4. If you were to write an online petition about language testing, what briefly would you say?



FURTHER READING

Ethical issues

Rea-Dickens (1997) considers the relationship between stakeholders in language testing and Hamp-Lyons (1997a) raises ethical concerns relating to backwash, impact and validity. These two papers form part of a special issue of *Language Testing* 14, 3 which is devoted to ethics in language testing. For an early discussion of the ethics of language testing, see Spolsky (1981). A. Brown (2012) discusses ethics in language testing and assessment. Boyd and Davies (2002) discuss issues in the development of codes of ethics and of practice. The International Language Testing Association (ILTA) has developed a Code of Ethics and Guidelines for Practice, both of which are to be found online and can be downloaded. Shohamy (2001) discusses the role of language tests within educational, social and political contexts. McNamara and Roever (2006) is an extensive treatment of the social dimensions of language testing.

Test impact

Gipps (1990) and Raven (1991) draw attention to the possible dangers of inappropriate assessment. Katz (2012) writes on the integration of assessment with teaching aims and learning. For an account of how the introduction of a new test can have a striking positive effect on teaching and learning, see Hughes (1988a).

Multiple measures

Benzehra (2018) provides an overview of multiple measures assessment. Chester (2005) presents a framework for combining multiple measures to reach high-stakes decisions.

Assessment literacy

Language Testing 30, 3 (2013) is a special issue on language assessment literacy. Taylor (2009) writes on the development of assessment literacy [ARAL 29, 21–36]. Ryan (2011) reviews three books on language testing and migration and citizenship. Shohamy and McNamara (2009) discuss the use of language tests for citizenship, immigration and asylum. Stansfield (2008) argues that language testers should become involved in public policy. Coombe et al. (2012c) discuss assessment literacy and make recommendations for its achievement. Lam (2015) points to a lack of language assessment literacy in Hong Kong and makes recommendations for improving the situation.

Attitudes of test-takers

Huhta et al. (2006) report on a longitudinal study of high school students' attitudes to a high-stakes test, using oral diaries.

2 Testing as problem solving: an overview of the book

Language testers are sometimes asked to say what is 'the best test' or 'the best testing technique'. Such questions reveal a misunderstanding of what is involved in the practice of language testing. A test that proves ideal for one purpose may be quite useless for another; a technique that may work very well in one situation can be entirely inappropriate in another. What suits large testing corporations may be quite out of place in the tests of teaching institutions. Equally, two teaching institutions may require very different tests, depending on the objectives of their courses, the purpose of the tests, and the resources available. Each testing situation is unique and sets a particular testing problem. And so the first step must be to state this testing problem as clearly as possible. Whatever test or testing system we then create should be one that:

- consistently provides accurate measures of precisely the abilities¹ in which we are interested;
- has a positive influence on teaching (in those cases where the test is likely to influence teaching);
- is economical in terms of time and money.

The first thing that testers have to be clear about is the purpose of testing in any particular situation. Different purposes will usually require different kinds of tests. This may seem obvious but it is something that is not always recognised. The purposes of testing discussed in this book are:

- To measure language proficiency.
- To discover how successful students have been in achieving the objectives of a course of study.
- To diagnose students' strengths and weaknesses, to identify what they know and what they don't know.
- To assist placement of students by identifying the stage or part of a teaching programme most appropriate to their ability.

¹'Abilities' is not being used here in any technical sense. It refers simply to what people can do in, or with, a language. It could, for example, include the ability to converse fluently in a language, as well as the ability to recite grammatical rules (if that is something which we are interested in measuring!). It does not, however, refer to language aptitude, the talent which people have, in differing degrees, for learning languages. The measurement of this talent in order to predict how well or how quickly individuals will learn a foreign language, is beyond the scope of this book. The interested reader is referred to Wen et al. (2019).

All of these purposes are discussed in the next chapter. That chapter also introduces different kinds of testing and test techniques: direct as opposed to indirect testing; discrete-point versus integrative testing; criterion-referenced testing as against norm-referenced testing; objective and subjective testing; paper-and-pencil tests versus computer-based tests; communicative language testing.

In stating the testing problem in general terms above, we spoke of providing consistent measures of precisely the abilities we are interested in. A test that does this is said to be *valid*. Chapter 4 addresses itself to various kinds of validity. It provides advice on the achievement of validity in test construction and shows how validity is measured.

The word 'consistently' was used in the statement of the testing problem. The consistency with which accurate measurements are made is in fact an essential ingredient of validity. If a test measures consistently (if, for example, a person's score on the test is likely to be very similar regardless of whether they happen to take it on, say, Monday morning rather than on Tuesday afternoon, assuming that there has been no significant change in their ability), it is said to be reliable. Reliability, already referred to in the previous chapter, is an absolutely essential quality of tests – what use is a test if it will give widely differing estimates of an individual's (unchanged) ability? – yet it is something which is distinctly lacking in too many teacher-made tests. Chapter 5 gives advice on how to achieve reliability and explains how it can be measured.

The concept of backwash was introduced in the previous chapter. Chapter 6 identifies a number of conditions for tests to meet in order to achieve positive backwash.

All tests cost time and money – to prepare, administer, score and interpret. As both are in limited supply, there is often likely to be a conflict between what appears to be a perfect testing solution in a particular situation and considerations of practicality. This issue is also discussed in Chapter 6.

The second half of the book is devoted to more detailed advice on the construction and use of tests – the putting into practice of the principles outlined in earlier chapters. Chapter 7 outlines and exemplifies the various stages of test development. Chapter 8 discusses a number of common testing techniques. Chapters 9–13 show how a variety of language abilities can best be tested, particularly within teaching institutions. Chapter 14 discusses 'overall ability' and how it may be measured. Chapter 15 considers the particular problems that have to be faced when young learners are tested. Chapter 16 looks at ways other than testing by which to assess students' ability. Chapter 17 examines the influence new technology has already had on language testing and attempts to anticipate its future effects. Chapter 18 gives practical advice on the administration of tests.

We have to say something about statistics. Some understanding of statistics is useful, indeed necessary, for a proper appreciation of testing matters and for successful problem solving. In the chapters on validity and reliability, simple statistical notions are presented in terms that it is hoped everyone should be able to grasp. Chapter 19 deals in some detail with the statistical analysis of test results. Even here, however, the emphasis is on interpretation rather than on calculation. In fact, given the computing power and statistics software that is readily available these days, there is no real need for any calculation on the part of language testers. They simply need to understand the output of the computer programs which they (or others) use. Chapter 19 attempts to develop this understanding and, just as important, show how valuable statistical information can be in developing better tests.

Appendix 1 deals with the construction of item banks, in which items can be stored with associated information, including the results of the statistical analysis described in Chapter 19.

Appendix 2 is a checklist for teachers to consult when they are considering adopting a test or recommending one for their students to take.

3

Kinds of tests and testing

Tests and testing can be classified according to their purpose and the various features which they incorporate.

Test purposes

We begin by considering the purposes for which language testing is carried out: to measure proficiency, to measure achievement, to diagnose linguistic strengths and weaknesses, and to help place students in appropriate classes.

Proficiency tests

Proficiency tests are designed to measure people's ability in a language, regardless of any training they may have had in that language. The content of a proficiency test, therefore, is not based on the content or objectives of language courses that people taking the test may have followed. Rather, it is based on a specification of what candidates have to be able to do in the language in order to be considered proficient. This raises the question of what we mean by the word *proficient*.

In the case of some proficiency tests, 'proficient' means having sufficient command of the language for a particular purpose. One example would be a test used to determine whether a student's English is good enough to follow a course of study at a British university. Such a test may even attempt to take into account the level and kind of English needed to follow courses in particular subject areas. It might, for instance, have one form of the test for arts subjects, another for sciences, and so on. Other examples would be tests designed to discover whether someone can function successfully as a United Nations translator, or as an air traffic controller. One thing such tests have in common is that they attempt to measure language ability for a more or less specific purpose. Whatever the specific purpose to which the language is to be put, this will be reflected in the specification of test content at an early stage of a test's development. (Tests are often referred to as being for a specific purpose, for educational or academic purposes, for medical professionals. See Further reading for examples.)

There are other proficiency tests which, by contrast, do not have any occupation or course of study in mind. For them the concept of proficiency is more general. British examples of these would be Cambridge Assessment English's *B2 First* exam (previously known as *Cambridge First Certificate in English*

examination (FCE) and their C2 Proficiency exam (previously the *Cambridge Certificate of Proficiency in English examination*). The function of such tests is to discover whether candidates have reached a certain standard with respect to a set of specified abilities. The examining bodies responsible for such tests are independent of teaching institutions and so can be relied on by potential employers, etc. to make fair comparisons between candidates from different institutions and different countries. Proficiency tests should have detailed specifications saying just what it is that successful candidates have demonstrated that they can do. Each test should be seen to be based directly on these specifications. All users of a test (teachers, students, employers, etc.) can then judge whether the test is suitable for them, and can interpret test results. It is not enough to have some vague notion of proficiency, however prestigious the testing body concerned. The Cambridge examinations referred to above are linked to the *Common European Framework of Reference (CEFR)*, B2 and C2 being levels in that framework (see Further reading).

Despite differences between them of content and level of difficulty, all proficiency tests have in common the fact that they are not based on courses that candidates may have previously taken. On the other hand, as we saw in Chapter 1, such tests may themselves exercise considerable influence over the method and content of language courses. Their backwash effect – for this is what it is – may be positive or harmful. In our view, the effect of some widely used proficiency tests has been more harmful than positive. However, the teachers of students who take such tests, and whose work suffers from a harmful backwash effect, may be able to exercise more influence over the testing organisations concerned than they realise. The supplementing of *TOEFL*® with a writing test, referred to in Chapter 1, is a case in point.

Achievement tests

Most teachers are unlikely to be responsible for proficiency tests. It is much more probable that they will be involved in the preparation and use of achievement tests. In contrast to proficiency tests, achievement tests are directly related to language courses, their purpose being to establish how successful individual students, groups of students, or the courses themselves have been in achieving objectives. They are of two kinds: final achievement tests and progress achievement tests.

Final achievement tests are those administered at the end of a course of study. They may be written and administered by ministries of education, official examining boards, or by members of teaching institutions. Clearly the content of these tests must be related to the courses with which they are concerned, but the nature of this relationship is a matter of disagreement amongst language testers.

In the view of some testers, the content of a final achievement test should be based directly on a detailed course syllabus or on the books and other

materials used. This has been referred to as the *syllabus-content approach*. It has an obvious appeal, since the test only contains what it is thought that the students have actually encountered, and thus can be considered, in this respect at least, a fair test. The disadvantage is that if the syllabus is badly designed, or the books and other materials are badly chosen, the results of a test can be very misleading. Successful performance on the test may not truly indicate successful achievement of course objectives. For example, a course may have as an objective the development of conversational ability, but the course itself and the test may require students only to utter carefully prepared statements about their home town, the weather, or whatever. Another course may aim to develop a reading ability in German, but the test may limit itself to the vocabulary the students are known to have met. Yet another course is intended to prepare students for university study in English, but the syllabus (and so the course and the test) may not include listening (with note-taking) to English delivered in lecture style on topics of the kind that the students will have to deal with at university. In each of these examples – all of them based on actual cases – test results will fail to show what students have achieved in terms of course objectives.

The alternative approach is to base the test content directly on the objectives of the course. This has a number of advantages. First, it compels course designers to be explicit about objectives. Secondly, it makes it possible for performance on the test to show just how far students have achieved those objectives. This in turn puts pressure on those responsible for the syllabus and for the selection of books and materials to ensure that these are consistent with the course objectives. Tests based on objectives work against the perpetuation of poor teaching practice, something which course-content-based tests, almost as if part of a conspiracy, fail to do. It is our belief that to base test content on course objectives is much to be preferred; it will provide more accurate information about individual and group achievement, and it is likely to promote a more positive backwash effect on teaching¹.

Now it might be argued that to base test content on objectives rather than on course content is unfair to students. If the course content does not fit well with objectives, they will be expected to do things for which they have not been prepared. In a sense this is true. But in another sense it is not. If a test is based on the content of a poor or inappropriate course, the students taking it will be misled as to the extent of their achievement and the quality of the course. Whereas if the test is based on objectives, not only will the information it gives be more useful, but there is also less chance of the course surviving in

¹ Of course, if objectives are unrealistic, then tests will also reveal a failure to achieve them. This, too, can only be regarded as salutary. There may be disagreement as to why there has been a failure to achieve the objectives, but at least this provides a starting point for necessary discussion which otherwise might never have taken place.

its present unsatisfactory form. Initially some students may suffer, but future students will benefit from the pressure for change. The long-term interests of students are best served by final achievement tests whose content is based on course objectives.

One encouraging recent development has been the increasing use of the *CEFR* (referred to above) in the writing of test specifications, in syllabus design, and in determining course book content. For example, the *CEFR* 'Can-do' descriptors (statements of what successful learners can do in a language at a particular level) are particularly useful in establishing objectives for learning and elements to be tested. For example, in reading: 'At the lowest level of reading ability, the learner can understand basic notices, instructions or information'.

At the next higher level: 'a learner can understand straightforward information within a known area, such as on products and signs and simple textbooks or reports on familiar matters'. The use of the *CEFR* is discussed further in Chapter 13.

The reader may wonder at this stage whether there is any real difference between final achievement tests and proficiency tests. If a test is based on the objectives of a course, and these are equivalent to the language needs on which a proficiency test is based, there is no reason to expect a difference between the form and content of the two tests. Two things have to be remembered, however. First, objectives and needs will not typically coincide in this way. Secondly, many achievement tests are not in fact based on course objectives. These facts have implications both for the users of test results and for test writers. Test users have to know on what basis an achievement test has been constructed, and be aware of the possibly limited validity and applicability of test scores. Test writers, on the other hand, must create achievement tests that reflect the objectives of a particular course, and not expect a general proficiency test (or some imitation of it) to provide a satisfactory alternative.

Progress achievement tests, as their name suggests, are intended to measure the progress that students are making. Since 'progress' is towards the achievement of course objectives, these tests, too, should relate to objectives. But how? One way of measuring progress would be repeatedly to administer final achievement tests, the (hopefully) increasing scores indicating the progress made. This is not really feasible, particularly in the early stages of a course. The low scores obtained would be discouraging to students and quite possibly to their teachers. The alternative is to establish a series of well-defined short-term objectives. These should make a clear progression towards the final achievement test based on course objectives. Then if the syllabus and teaching are appropriate to these objectives, progress tests based on short-term objectives will fit well with what has been taught. If not, there will be pressure to create a better fit. If it is the syllabus that is at fault, it is the tester's responsibility to make clear that it is there that change is needed, not in the tests.

In addition to more formal achievement tests that require careful preparation, teachers should feel free to set their own 'pop quizzes'. These serve both to make a rough check on students' progress and to keep students on their toes. Since such tests will not form part of formal assessment procedures, their construction and scoring need not be too rigorous. Nevertheless, they should be seen as measuring progress towards the intermediate objectives on which the more formal progress achievement tests are based. They can, however, reflect the particular 'route' that an individual teacher is taking towards the achievement of objectives.

It has been argued in this section that it is better to base the content of achievement tests on course objectives rather than on the detailed content of a course. However, it may not be at all easy to convince colleagues of this, especially if the latter approach is already being followed. Not only is there likely to be natural resistance to change, but such a change may represent a threat to many people. A great deal of skill, tact and, possibly, political manoeuvring may be called for – topics on which this book cannot pretend to give advice.

This is an appropriate moment for us to make the distinction between *summative assessment* and *formative assessment*. Summative assessment is designed to measure the outcome of a period of instruction. Achievement tests normally form part of summative assessment. In fact, they are often the only component of such assessment, something we will discuss in the next chapter.

Formative assessment, on the other hand, is designed to help students assess their own learning, and assist instructors to identify students facing problems, and modify the instruction accordingly. It is carried out informally on a day-to-day basis and provides the students with feedback on their control of what is being taught. Chapter 16 includes advice on formative assessment.

Diagnostic tests

Diagnostic tests are used to identify learners' strengths and weaknesses. They are intended primarily to ascertain what learning still needs to take place. At the level of broad language skills this is reasonably straightforward. We can be fairly confident of our ability to create tests that will tell us that someone is particularly weak in, say, speaking as opposed to reading in a language. Indeed existing proficiency tests may often prove adequate for this purpose.

We may be able to go further, and analyse samples of a person's performance in writing or speaking in order to create profiles of their ability with respect to such categories as 'grammatical accuracy' or 'linguistic appropriacy'. Indeed Chapters 9 and 10 suggest that raters of writing and speaking test performance should provide feedback to the test-takers as a matter of course.

But it is not so easy to obtain a detailed analysis of a student's command of grammatical structures – something that would tell us, for example,

whether she or he had mastered the present perfect/past simple distinction in English. In order to be sure of this, we would need a number of examples of the choice the student made between the two structures in every different context that we thought was significantly different and important enough to warrant obtaining information on. A single example of each would not be enough, since a student might give the correct response by chance. Similarly, if one wanted to test control of the English article system, one would need several items for each of the twenty or so uses of the articles (including the 'zero article') listed in *Collins Cobuild English Usage* (1992). What is more, we would probably wish to include items that tested the student's productive ability, as well as others that tested their receptive ability. Thus, a comprehensive diagnostic test of English grammar would be vast (think of what would be involved in testing the modal verbs, for instance). The size of such a test would make it impractical to administer in a routine fashion. For this reason, very few tests are constructed for purely diagnostic purposes, and those that there are tend not to provide very detailed or reliable information. One diagnostic test which deserves attention, though its output is not very detailed, is *DIALANG*, which offers versions in fourteen European languages, each having five modules: reading, writing, listening, grammatical structures, and vocabulary.

The lack of good, detailed diagnostic tests is unfortunate. They could be extremely useful for individualised instruction or self-instruction. Learners would be shown where gaps exist in their command of the language, and could be directed to sources of information, exemplification and practice. In the previous edition of this book, the hope was expressed that the ready availability of powerful but relatively inexpensive computers with very large memories would change the situation. Well-written computer programs, it was suggested, would ensure that the learner spends no more time than is absolutely necessary to obtain the desired information, and without the need for a test administrator. However, it was admitted that whether or not they became generally available would depend on the willingness of individuals to write them and of publishers to distribute them. Unfortunately, at the time of writing, neither publishers nor testing organisations appear so far to have thought the necessary investment of time and money to be worth their while.

Placement tests

Placement tests, as their name suggests, are intended to provide information that will help to place students at the stage (or in the part) of the teaching programme most appropriate to their abilities. Typically they are used to assign students to classes at different levels. Placement tests can be bought, but this is to be recommended only when the institution concerned is sure that the test being considered suits its particular teaching programme. No one placement test will work for every institution, and the initial assumption about any test that is commercially available must be that it will not work well. One possible exception is placement tests designed for use by language schools, where the similarity of popular textbooks used in

them means that the schools' teaching programmes also tend to resemble each other.

The placement tests that are most successful are those constructed for particular situations. They depend on the identification of the key features at different levels of teaching in the institution. They are tailor-made rather than bought off the peg. This usually means that they have been produced 'in house'. The work that goes into their construction is rewarded by the saving in time and effort through accurate placement. An example of how a placement test might be developed is given in Chapter 7; the validation of placement tests is referred to in Chapter 4.

It is worth adding, perhaps, that too much should not be expected of a placement test. Where feasible, the test would benefit from being supplemented by a brief interview. The student's gender, age, nationality, personality and motivation, and other factors, are likely to affect how well they are suited to a particular class (Green 2012). As with other kinds of assessment, there has to be a trade-off between accuracy and efficiency. In placement testing, since errors can be relatively easily rectified, accuracy is less important than in high-stakes tests.

Screening tests

Screening tests are used in order to avoid the expense and loss of time taken to administer longer, more complex tests when they are not necessary. An example from our own experience was at an English-medium university overseas, when a lengthy high-stakes proficiency test effectively determined whether incoming students could proceed directly to their undergraduate studies or had to spend time studying English full-time for up to a year. Since from experience it was known that most new students would fail the proficiency test, a straightforward multiple choice was given first. The cut-off point for this screening test was set at a level that allowed us to be confident that students who did not reach it could not possibly pass the proficiency test. Students who did score at or above the cut-off point were allowed to take the proficiency test.

Test Features

So far in this chapter we have classified tests according to their purpose. We now go on to look at contrasting features of test construction.

Direct versus indirect testing

Testing is said to be *direct* when it requires the candidate to perform precisely the skill that we wish to measure. If we want to know how well candidates can write compositions, we get them to write compositions. If we want to know how well they pronounce a language, we get them to speak. The tasks, and the texts that are used in direct testing, should be as authentic as possible. The fact that candidates are aware that they are in a

test situation means that the tasks cannot be really authentic. Nevertheless every effort is made to make them as realistic as possible.

Direct testing is easier to carry out when it is intended to measure the productive skills of speaking and writing. The very acts of speaking and writing provide us with information about the candidate's ability. With listening and reading, however, it is necessary to get candidates not only to listen or read but also to demonstrate that they have done this successfully. Testers have to devise methods of eliciting such evidence accurately and without the method interfering with the performance of the skills in which they are interested. Appropriate methods for achieving this are discussed in Chapters 11 and 12. Interestingly enough, in many texts on language testing it is the testing of productive skills that is presented as being most problematic, for reasons usually connected with reliability. In fact these reliability problems are by no means insurmountable, as we shall see in Chapters 9 and 10.

Direct testing has a number of attractions. First, provided that we are clear about just what abilities we want to assess, it is relatively straightforward to create the conditions which will elicit the behaviour on which to base our judgements. Secondly, at least in the case of the productive skills, the assessment and interpretation of students' performance is also quite straightforward. Thirdly, since practice for the test involves practice of the skills that we wish to foster, there is likely to be a helpful backwash effect.

Indirect testing attempts to measure the abilities that underlie the skills in which we are interested. There was a time when some professional testers would use the multiple choice technique to measure writing ability. Their items were of the following kind where the candidate had to identify which of the underlined elements is erroneous or inappropriate in formal standard English:

At the outset the judge seemed unwilling to believe anything that was said to her by my wife and I.

While the ability to respond to such items has been shown to be related statistically to the ability to write compositions (although the strength of the relationship was not particularly great), the two abilities are far from being identical. Another example of indirect testing is Lado's (1961) proposed method of testing pronunciation ability by a paper-and-pencil test in which the candidate has to identify pairs of words which rhyme with each other.

Perhaps the main appeal of indirect testing is that it seems to offer the possibility of testing a representative sample of a finite number of abilities which underlie a potentially indefinite large number of manifestations of them. If, for example, we take a representative sample of grammatical structures, then, it may be argued, we have taken a sample which is relevant for all the situations in which control of grammar is necessary. By contrast, direct testing is inevitably limited to a rather small sample of tasks, which may call on a restricted and possibly unrepresentative range

of grammatical structures. On this argument, indirect testing is superior to direct testing in that its results are more generalisable.

The main problem with indirect tests is that the relationship between performance on them and performance of the skills in which we are usually more interested tends to be rather weak in strength and uncertain in nature. We do not yet know enough about the component parts of, say, composition writing to predict accurately composition writing ability from scores on tests that measure the abilities that we believe underlie it. We may construct tests of grammar, vocabulary, discourse markers, handwriting, punctuation, or of any other linguistic element. But we will still not be able to predict accurately scores on compositions (even if we make sure of the validity of the composition scores by having people write many compositions and by scoring these in a valid and highly reliable way).

It seems to us that in our present state of knowledge, at least as far as proficiency and final achievement tests are concerned, it is preferable to rely principally on direct testing. Provided that we sample reasonably widely (for example require at least two compositions, each calling for a different kind of writing and on a different topic), we can expect more accurate estimates of the abilities that really concern us than would be obtained through indirect testing. The fact that direct tests are generally easier to construct simply reinforces this view with respect to institutional tests, as does their greater potential for positive backwash. It is only fair to say, however, that many testers are reluctant to commit themselves entirely to direct testing and will always include an indirect element in their tests. Of course, to obtain diagnostic information on underlying abilities, such as control of particular grammatical structures, indirect testing may be perfectly appropriate.

In summary, we might say that both direct and indirect testing rely on obtaining samples of behaviour and drawing inferences from them. While sampling may be easier in indirect testing, making meaningful inferences is likely to be more difficult. Accurate inferences may be more readily made in direct testing, though it may be more difficult to obtain samples that are truly representative. One can expect the backwash effect of direct testing to be the more positive.

Before ending this section, it should be mentioned that some tests are referred to as *semi-direct*. The most obvious examples of these are speaking tests where candidates respond to recorded stimuli, with their own responses being recorded and later scored. These tests are semi-direct in the sense that, although not direct, they simulate direct testing.

Discrete point versus integrative testing

Discrete point testing refers to the testing of one element at a time, item by item. This might, for example, take the form of a series of items, each testing a particular grammatical structure. *Integrative testing*, by contrast,

requires the candidate to combine many language elements in the completion of a task. This might involve writing a composition, making notes while listening to a lecture, taking a dictation, or completing a cloze passage. Clearly this distinction is not unrelated to that between indirect and direct testing. Discrete point tests will almost always be indirect, while integrative tests will tend to be direct. However, some integrative testing methods, such as the cloze procedure, are indirect. Diagnostic tests of grammar of the kind referred to in an earlier section of this chapter will tend to be discrete point.

Norm-referenced versus criterion-referenced testing

Imagine that a reading test is administered to an individual student. When we ask how the student performed on the test, we may be given two kinds of answer. An answer of the first kind would be that the student obtained a score that placed her or him in the top 10 percent of candidates who have taken that test, or in the bottom five percent; or that she or he did better than 60 percent of those who took it. A test which is designed to give this kind of information is said to be *norm-referenced*. It relates one candidate's performance to that of other candidates. We are not told directly what the student is capable of doing in the language.

The other kind of answer we might be given is exemplified by the following, taken from the Interagency Language Roundtable (ILR) language skill level descriptions for reading:

R-3: Reading 3 (General Professional Proficiency) *Able to read within a normal range of speed and with almost complete comprehension a variety of authentic prose material on unfamiliar subjects. Reading ability is not dependent on subject matter knowledge, although it is not expected that the individual can comprehend thoroughly subject matter which is highly dependent on cultural knowledge or which is outside his/her general experience and not accompanied by explanation. Text-types include news stories similar to wire service reports or international news items in major periodicals, routine correspondence, general reports, and technical material in his/her professional field; all of these may include hypothesis, argumentation and supported opinions. Misreading rare. Almost always able to interpret material correctly, relate ideas and "read between the lines," (that is, understand the writers' implicit intents in text of the above types). Can get the gist of more sophisticated texts, but may be unable to detect or understand subtlety and nuance. Rarely has to pause over or reread general vocabulary. However, may experience some difficulty with unusually complex structure and low frequency idioms.*

Testing for assignment to levels is intended to be carried out in a face-to-face situation, with questions being asked orally. The tester gives the candidate reading matter of different kinds and at different levels of difficulty, until a conclusion can be made as to the candidate's ability. This can only be done, of course, with relatively small numbers of candidates.

In this case we learn nothing about how the individual's performance compares with that of other candidates. Rather we learn something about what he or she can actually do in the language. Tests that are designed to provide this kind of information directly are said to be *criterion-referenced*².

When the previous edition of this book was published, it was not difficult to point to major language tests which were norm-referenced. The scores which were reported did not indicate what a candidate could or could not do. Rather a numerical score was provided, which candidates, teachers and institutions had to interpret on the basis of experience. Only over time did it become possible to relate a person's score to their likely success in coping in particular second or foreign language situations.

This is no longer the case. More typical now is *IELTS* (*International English Language Testing System*) of the British Council, Cambridge Assessment English and the University of Cambridge, which is described on a British Council website as criterion-referenced. On the basis of their performance on the test, candidates are given a band score (or a 'half band' score of, for example, 6.5, for a candidate falling between two 'full bands'). The bands are:

Band score	Skill level	Description
9	Expert user	The test taker has fully operational command of the language. Their use of English is appropriate, accurate and fluent, and shows complete understanding.
8	Very good user	The test taker has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriate usage. They may misunderstand some things in unfamiliar situations. They handle complex and detailed argumentation well.
7	Good user	The test taker has operational command of the language, though with occasional inaccuracies, inappropriate usage and misunderstandings in some situations. They generally handle complex language well and understand detailed reasoning.

² People differ somewhat in their use of the term 'criterion-referenced'. This is unimportant provided that the sense intended is made clear. The sense in which it is used here is the one which we feel will be most useful to the reader in analysing testing problems.

Band score	Skill level	Description
6	Competent user	The test taker has an effective command of the language despite some inaccuracies, inappropriate usage and misunderstandings. They can use and understand fairly complex language, particularly in familiar situations.
5	Modest user	The test taker has a partial command of the language and copes with overall meaning in most situations, although they are likely to make many mistakes. They should be able to handle basic communication in their own field.
4	Limited user	The test taker's basic competence is limited to familiar situations. They frequently show problems in understanding and expression. They are not able to use complex language.
3	Extremely limited user	The test taker conveys and understands only general meaning in very familiar situations. There are frequent breakdowns in communication.
2	Intermittent user	The test taker has great difficulty understanding spoken and written English.
1	Non-user	The test taker has no ability to use the language except a few isolated words.

(Adapted from Cambridge Assessment English)

A test that assigns candidates to bands in this fashion does indeed seem to be criterion-referenced. The *IELTS* descriptors for speaking offer confirmation. Those for Band 7, for example, are:

Fluency and Coherence:

- speaks at length without noticeable effort or loss of coherence
- may demonstrate language related hesitation at times, or some repetition and/or self-correction
- uses a range of connectives and discourse markers with some flexibility

Lexical Resource:

- uses vocabulary resource flexibly to discuss a variety of topics
- uses some less common and idiomatic vocabulary and shows some awareness of style and collocation, with some inappropriate choices
- uses paraphrase effectively

Grammatical Resource:

- uses a range of complex structures with some flexibility
- frequently produces error-free sentences, though some grammatical mistakes persist
- shows all the positive features of Band 6 and some, but not all, of the positive features of Band 8
- is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation

- uses a range of connectives and discourse markers but not always appropriately
- has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies
- generally paraphrases successfully
- uses a mix of simple and complex structures, but with limited flexibility
- may make frequent mistakes with complex structures, though these rarely cause comprehension problems

Pronunciation

- uses a range of pronunciation features with mixed control
- shows some effective use of features but this is not sustained
- can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times

(*IELTS SPEAKING*: Band Descriptors, public version)

However, unlike ILR, *IELTS* does not have a set of descriptors for Reading, even though it assigns candidates to a Reading band. It does this on the basis of performance on a set of 40 items, because face-to-face, adaptive testing of reading is not feasible for the tens of thousands taking *IELTS*.

Scores on the Reading section have to be converted into Band scores by means of statistical manipulation which make use of items whose difficulty level is known before the test is taken. A description of this process is beyond the compass of this book, but see Chapter 19 and Appendix 1 to have some idea of what it involves. However, approximate Band scores can be calculated using the following table.

Indicative <i>IELTS</i> score/band transformation table (Academic Reading)		
Band	Min Score	Max Score
5	13	17
5.5	18	21
6	22	25
6.5	26	29
7	30	32
7.5	33	34
8	35	36

The indirectness between performance on parts of *IELTS* and the Band scores to which candidates are assigned brings into question *IELTS*'s status as a criterion-referenced test. Saville, Director, Research and Thought Leadership, Cambridge Assessment English, believes that it is criterion-referenced but that it is 'the weak variant of criterion-referenced testing' (personal communication 2018). We are prepared to accept this meaning for *IELTS* overall, but the sense intended is not 'pure' criterion-referencing as we understand it.

Pure criterion-referenced tests classify people according to whether or not they are able to perform some task or set of tasks satisfactorily. The tasks

are set, and the performances are evaluated. It does not matter in principle whether all the candidates are successful, or none of the candidates is successful. In broad terms, tasks are set, and those who perform them satisfactorily 'pass'; those who don't, 'fail'. This means that students are encouraged to measure their progress in relation to meaningful criteria, without feeling that, because they are less able than most of their fellows, they are destined to fail. Criterion-referenced tests therefore have two positive virtues: they set meaningful standards in terms of what people can do, which do not change with different groups of candidates, and they motivate students to attain those standards. We welcome the trend to make major tests more criterion-referenced.

Books on language testing have tended to give advice which is more appropriate to norm-referenced testing than to criterion-referenced testing. One reason for this may be that procedures for use with norm-referenced tests (particularly with respect to such matters as the analysis of items and the estimation of reliability) are well established, while those for criterion-referenced tests are not. The view taken in this book, and argued for in Chapter 6, is that criterion-referenced tests are often to be preferred, not least for the positive backwash effect they are likely to have. The lack of agreed procedures for such tests is not sufficient reason for them to be excluded from consideration. Chapter 5 presents one method of estimating the consistency (more or less equivalent to 'reliability') of criterion-referenced tests.

Objective testing versus subjective testing

The distinction here is between methods of scoring, and nothing else. If no judgement is required on the part of the scorer, then the scoring is *objective*. A multiple choice test, with the correct responses unambiguously identified, would be a case in point. If judgement is called for, the scoring is said to be *subjective*. There are different degrees of subjectivity in testing. The impressionistic scoring of a composition may be considered more subjective than the scoring of short answers in response to questions on a reading passage.

Objectivity in scoring is sought after by many testers, not for itself, but for the greater reliability it brings. In general, the less subjective the scoring, the greater agreement there will be between two different scorers (and between the scores of one person scoring the same test paper on different occasions). However, there are ways of obtaining reliable subjective scoring, even of compositions. These are discussed first in Chapter 5.

Means of test delivery

Tests can be paper-and-pencil face-to-face or computer-based.

Paper-and-pencil tests

The traditional language test is printed on paper with the candidate responding with pen or pencil. One drawback of such tests is their lack of flexibility. The order of items is fixed, usually in ascending order of difficulty, and candidates are required to respond to all of them. This is not the most economical way of collecting information on someone's ability. People of high ability (in relation to the test as a whole) will spend time responding to items that are very easy for them – all, or nearly all, of which they will get correct. We would have been able to predict their performance on these items from their correct response to more difficult items. Similarly, we could predict the performance of people of low ability on difficult items, simply by seeing their consistently incorrect response to easy items.

The other drawback is obvious. By themselves, paper-and-pencil tests cannot measure ability in the spoken language.

Face-to-face tests

A face-to-face test, in which one or more testers interact with one or more candidates, is clearly more flexible: the testers can adapt to the candidates' responses. It also allows the measurement of spoken ability. Its principal drawback is its cost in terms of time, effort and, when the testers are paid for their work, money. The cost has to be weighed against the value of the information obtained and the backwash effect of such testing. Most face-to-face testing is of speaking ability.

Computer-based tests

Computer-based tests can be on the internet, on an intranet, or on a stand-alone computer. One advantage of having a test on a computer is that it can be taken at any time, frequently without the need for supervision, and results can often be reported immediately. It also allows for testing to be adaptive. *Computer adaptive testing* offers a potentially more efficient way of collecting information on people's ability. All candidates are presented initially with an item of average difficulty. Those who respond correctly are presented with a more difficult item; those who respond incorrectly are presented with an easier item. The computer goes on in this way to present individual candidates with items that are appropriate for their apparent level of ability (as estimated by their performance on previous items), raising or lowering the level of difficulty until a dependable estimate of their ability is achieved. This dependable estimate, which will normally be arrived at after collecting responses to a relatively small number of items, is based on statistical analysis (item response theory) which most language teachers may find daunting but which is presented briefly in Chapter 19³.

³ The kind of diagnostic testing which we would like to see would also be computer adaptive.

One thing that computer-based tests cannot do satisfactorily is measure a candidate's ability to interact with another speaker, something which may be regarded as an essential component of speaking ability. Similarly, it is difficult to see how meaning can be taken into account in the scoring of written work when scoring is carried out solely by a computer program. That said, in recent years we have seen such remarkable advances in, for example, machine translation (which also involves meaning), that we should be wary of discounting any future possibilities.

One last point about online proficiency testing is that special care needs to be taken to ensure that security is maintained. In particular, the identity of the test-taker may be confirmed by, for example, digital signature or palm scanning.

Communicative language testing

Much has been written about 'communicative language testing'. Discussions have centred on the desirability of measuring the ability to take part in acts of communication (including reading and listening) and on the best way to do this. It is assumed in this book that it is usually communicative ability that we want to test. As a result, what we believe to be the most significant points made in discussions of communicative testing are to be found throughout. A recapitulation under a separate heading would therefore be redundant. As one of its first proponents wrote recently, 'perhaps the most interesting thing about the phrase "communicative language testing" is that it belongs very clearly to history' (Morrow 2012).



READER ACTIVITIES

Consider a number of language tests with which you are familiar. For each of them, answer the following questions:

1. What is the purpose of the test?
2. Does it represent direct or indirect testing (or a mixture of both)?
3. Are the items discrete point or integrative (or a mixture of both)?
4. Which items are objective, and which are subjective? Can you order the subjective items according to degree of subjectivity?
5. Is the test norm-referenced or criterion-referenced?
6. Does the test measure communicative abilities? Would you describe it as a communicative test? Justify your answers.
7. What relationship is there between the answers to question 6 and the answers to the other questions?
8. Would there be any difficulty in making any of the tests computer-based?

Take at least one module of *DIALANG* for a language that you know (not your first). Do the results seem to give an accurate account of your ability? Take the same modules again. Compare the two sets of results.

FURTHER READING

Achievement testing

For a discussion of the two approaches towards achievement test content specification, see Pilliner (1968).

Diagnostic testing

Nitko (2001) includes a chapter on diagnostic assessment. Alderson et al. (2015) draw lessons from other fields that may contribute to a theory of diagnosis in second and foreign language assessment. Alderson (2005) explores issues in diagnostic testing and provides information about the development of *DIALANG*. Green and Weir (2004) cast doubt on the feasibility of obtaining diagnostic information using a grammar-based placement test. Knoch (2009) compares two rating scales for the diagnosis of writing ability. Jang (2009) and Kim and Elder (2015) examine the possibility of carrying out diagnosis using non-diagnostic reading tests.

Placement testing

Language Testing 32, 3 (2015) is a special issue on the future of diagnostic language testing. Kokhan (2013) argues against using standardised test scores for placement on ESL courses. Wall et al. (1994), Fulcher (1997) and Green (2012) discuss issues in placement test development.

Indirect v. direct testing, authenticity

Direct testing calls for texts and tasks to be as authentic as possible: Volume 2, 1 (1985) of the journal *Language Testing* is devoted to articles on authenticity in language testing. *Language Testing* 33, 2 is devoted to the topic of authenticity in LSP (Language for Specific Purposes) testing. Lewkowicz (2000) discusses authenticity in language testing. A classic account of the development of an indirect test of writing is given in Godshalk et al. (1966).

Criterion-referenced testing v. norm-referenced testing

Hudson and Lynch (1984) was an early discussion of criterion-referenced language testing; Brown and Hudson's (2002) book is the first full-length treatment of the subject. Classic short papers on criterion-referencing and norm-referencing (not restricted to language testing) are by Popham (1978), favouring criterion-referenced testing, and Ebel (1978), arguing for the superiority of norm-referenced testing. Doubts about the applicability of criterion-referencing to language testing are expressed by Skehan (1984); for a different view, see Hughes (1986). Examples of criterion-referenced tests are: The ACTFL Oral Proficiency Interview (<http://www.actfl.org>); the FBI Listening summary translation exam (Scott et al. 1996); the Canadian Academic English Language (CAEL) Assessment (Jennings et al. 1999).

Discrete point v. integrative testing

Carroll (1961) made the distinction between discrete point and integrative language testing. Oller (1979) discusses integrative testing techniques.

Computer-based testing

Chalhoub-Deville and Deville (1999) look at computer adaptive language testing. Chalhoub-Deville (1999) edited a collection of papers discussing issues in computer adaptive testing of reading proficiency. Fulcher (2000) discusses the role of computers in language testing, as do Chapelle and Douglas (2006).

Communicative language testing

Morrow (1979) is a seminal paper on communicative language testing. Morrow (2012) is a more recent discussion of the topic. Further discussion of communicative language testing can be found in Canale and Swain (1980), Alderson and Hughes (1981, Part 1), Hughes and Porter (1983), and Davies (1988). Weir's (1990) book has as its title *Communicative Language Testing*.

Online resources

For examples of tests delivered via the internet, and information about them, we suggest that readers search for some or all of the following proficiency tests: *Pearson Test of English* (two versions: *Academic*, and *General*), *TOEFL*®, *TOEIC*®, *Oxford Test of English*.

Online placement tests include:

Cambridge English Placement Test, *Oxford Online Placement Test*, *Oxford Young Learners Placement Test*, *Pearson English Placement Test*, *The Dynamic Placement Test* (Clarity English).

Handbooks for the various Cambridge proficiency tests can be obtained online, as can information on *IELTS* and the Michigan test. Also available are descriptions for skills at various levels: ILR, ACTFL (the American Council for the Teaching of Foreign Languages), ALTE (Association of Language Testers in Europe, covering 25 languages). *DIALANG* can also be found online.

4

Validity

We already know from Chapter 2 that a test is said to be valid if it measures accurately what it is intended to measure. This makes validity the central concept in language testing¹.

What do we want language tests to measure? We want to measure essentially theoretical constructs such as 'reading ability', 'fluency in speaking', 'control of grammar', and so on. For this reason, the term *construct validity*² has come to be used to refer to the general, overarching notion of validity.

We try to create a test whose scores maximise the contribution of the construct in question and minimise the contribution of irrelevant factors (such as general knowledge, first language background, etc.)³.

However hard testers try to achieve this, it is not enough to assert that a test has construct validity; empirical evidence is needed. Such evidence may take several forms, including the subordinate forms of validity, *content validity* and *criterion-related validity*. We shall begin by looking at these two forms of evidence in turn, and attempt to show their relevance for the solution of language testing problems. We shall then turn to other forms of evidence of validity.

Content validity

The first form of evidence relates to the content of the test. A test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned. It is obvious that a grammar test, for instance, must be made up of items relating to the knowledge or control of grammar. But this in itself does not ensure content validity. The test would have content validity only if it included a proper sample of the relevant structures. Just what are

¹ Other testing practitioners would say that it is not the test itself, but test scores, or the uses to which a test is put, that are valid or not. These alternative views are examined briefly below.

² When the term 'construct validity' was first used, it was in the context of psychological tests, particularly of personality tests. There was real concern at that time at the number of such tests which purported to measure psychological constructs, without offering evidence that these constructs existed in a measurable form. The demand was therefore that such evidence of these constructs be provided as part of demonstrating a test's validity.

³ In the testing literature, this is often expressed as the need to minimise both 'construct under-representation' and 'construct irrelevant variance'.

the relevant structures will depend, of course, upon the purpose of the test. We would not expect an achievement test for intermediate learners to contain just the same set of structures as one for advanced learners. Similarly, the content of a reading test should reflect the particular reading skills (e.g., skimming for gist, scanning for information) and the type and difficulty of texts with which a successful candidate is expected to cope.

In order to judge whether or not a test has content validity, we need a *specification* of the skills or structures, etc. that it is meant to cover. Such a specification should be made at a very early stage in test construction. It isn't to be expected that everything in the specification will always appear in the test; there may simply be too many things for all of them to appear in a single test. But it will provide the test constructor with the basis for making a principled selection of elements for inclusion in the test. A comparison of test specification and test content is the basis for judgements as to content validity. Ideally these judgements should be made by people who are familiar with language teaching and testing but who are not directly concerned with the production of the test in question.

What is the importance of content validity? First, the greater a test's content validity, the more likely it is to be an accurate measure of what it is supposed to measure, i.e. to have construct validity. A test in which major areas identified in the specification are under-represented – or not represented at all – is unlikely to be accurate. Secondly, such a test is likely to have a harmful backwash effect. Areas that are not tested are likely to become areas ignored in teaching and learning. Too often the content of tests is determined by what is easy to test rather than what is important to test. The best safeguard against this is to write full test specifications and to ensure that the test content is a fair reflection of these. For this reason, content validation should be carried out while a test is being developed; it should not wait until the test is already being used. Where a test of language for a specific purpose is being designed, it is important to consult domain specialists (for example, air traffic controllers for a test of aviation English). Advice on the writing of specifications can be found in Chapter 7.

Criterion-related validity

The second form of evidence of a test's construct validity relates to the degree to which results on the test agree with those provided by some independent and highly dependable assessment of the candidate's ability. This independent assessment is thus the criterion measure against which the test is validated.

There are essentially two kinds of criterion-related validity: *concurrent validity* and *predictive validity*.

Concurrent validity

Concurrent validity is established when the test and the criterion are administered at about the same time. To exemplify this kind of validation in achievement testing, let us consider a situation where course objectives call for a speaking component as part of the final achievement test. The objectives may list a large number of 'functions' which students are expected to perform orally, to test all of which might take 45 minutes for each student. This could well be impractical. Perhaps it is felt that only ten minutes can be devoted to each student for the speaking component. The question then arises: can such a ten-minute session give a sufficiently accurate estimate of the student's ability with respect to the functions specified in the course objectives? Is it, in other words, a valid measure?

From the point of view of content validity, this will depend on how many of the functions are tested in the component, and how representative they are of the complete set of functions included in the objectives. Every effort should be made when designing the speaking component to give it content validity. Once this has been done, however, we can go further. We can attempt to establish the concurrent validity of the component.

To do this, we should choose at random a sample of all the students taking the test. These students would then be subjected to the full 45-minute speaking component necessary for coverage of all the functions, using perhaps four scorers to ensure reliable scoring (see Chapter 5). This would be the criterion test against which the shorter test would be judged. The students' scores on the full test would be compared with the ones they obtained on the ten-minute session, which would have been conducted and scored in the usual way, without knowledge of their performance on the longer version. If the comparison between the two sets of scores reveals a high level of agreement, then the shorter version of the speaking component may be considered valid, inasmuch as it gives results similar to those obtained with the longer version. If, on the other hand, the two sets of scores show little agreement, the shorter version cannot be considered valid; it cannot be used as a dependable measure of achievement with respect to the functions specified in the objectives. Of course, if ten minutes really is all that can be spared for each student, then the speaking component may be included for the contribution that it makes to the assessment of students' overall achievement and for its backwash effect. But it cannot be regarded as an accurate measure in itself.

References to 'a high level of agreement' and 'little agreement' raise the question of how the level of agreement is measured. There are, in fact, standard procedures for comparing sets of scores in this way, which generate what is called a 'correlation coefficient' (or, when we are considering validity, a 'validity coefficient') – a mathematical measure of similarity. Perfect agreement between two sets of scores will result in a coefficient of 1. Total lack of agreement will give a coefficient of zero. To get a feel for the meaning of a coefficient between these two extremes, read the contents of the box on page 32.

Whether or not a particular level of agreement is regarded as satisfactory will depend upon the purpose of the test and the importance of the decisions that are made on the basis of it. If, for example, a test of oral ability was to be used as part of the selection procedure for a high-level diplomatic post, then a coefficient of 0.7 might well be regarded as too low for a shorter test to be substituted for a full and thorough test of oral ability. The saving in time would not be worth the risk of appointing someone with insufficient ability in the relevant foreign language. On the other hand, a coefficient of the same size might be perfectly acceptable for a brief interview forming part of a placement test⁴.



UNDERSTANDING LEVELS OF AGREEMENT

To get a feel for what a coefficient means in terms of the level of agreement between two sets of scores, it is best to square that coefficient. Let us imagine that a coefficient of 0.7 is calculated between the two oral tests referred to in the main text. Squared, this becomes 0.49. If this is regarded as a proportion of one, and converted to a percentage, we get 49 percent. On the basis of this, we can say that the scores on the short test predict 49 percent of the variation in scores on the longer test. In broad terms, there is almost a 50 percent agreement between one set of scores and the other. A coefficient of 0.5 would signify 25 percent agreement; a coefficient of 0.8 would indicate 64 percent agreement. It is important to note that a 'level of agreement' of, say, 50 percent does not mean that 50 percent of the students would each have equivalent scores on the two versions. We are dealing with an overall measure of agreement that does not refer to the individual scores of students. This explanation of how to interpret validity coefficients is very brief and necessarily rather crude. For a better understanding, the reader is referred to the Further reading section at the end of the chapter. Note that a perfect correlation would be 1.0. This would mean that one set of test scores would perfectly predict another set of scores, something which we cannot expect to occur in practice.

It should be said that the criterion for concurrent validation is not necessarily a proven, longer test. A test may be validated against, for example, teachers' assessments of their students, provided that the assessments themselves can be relied on. This would be appropriate where a test was developed that claimed to be measuring something different from all existing tests.

Predictive validity

The second kind of criterion-related validity is predictive validity. This concerns the degree to which a test can predict candidates' future performance. An example would be how well a proficiency test could predict a student's ability to cope with a graduate course at a British

⁴ Sometimes the size of a correlation coefficient can be misleading, an accident of the particular sample of people taking the test(s). If, for example, there are 'extreme' scores from outstandingly good or outstandingly poor takers of the test(s), the coefficient may be higher than the performance of the group as a whole warrants. See Nitko (2001) for details.

university. The criterion measure here might be an assessment of the student's English as perceived by his or her supervisor at the university, or it could be the outcome of the course (pass/fail etc.). The choice of criterion measure raises interesting issues. Should we rely on the subjective and untrained judgements of supervisors? How helpful is it to use final outcome as the criterion measure when so many factors other than ability in English (such as subject knowledge, intelligence, motivation, health and happiness) will have contributed to every outcome? Where outcome is used as the criterion measure, a validity coefficient of around 0.4 (less than 20 percent agreement) is about as high as one can expect. This is partly because of the other factors, and partly because those students whose English the test predicted would be inadequate are not normally permitted to take the course, and so the test's (possible) accuracy in predicting problems for those students goes unrecognised⁵.

As a result, a validity coefficient of this order is generally regarded as satisfactory. The Further reading section at the end of the chapter gives references to the reports on the validation of the British Council's *ELTS* test (the predecessor of *IELTS*), in which these issues are discussed at length.

Another example of predictive validity would be where an attempt was made to validate a placement test. Placement tests attempt to predict the most appropriate class for any particular student. Validation would involve an enquiry, once courses were under way, into the proportion of students who were thought to be misplaced. It would then be a matter of comparing the number of misplacements (and their effect on teaching and learning) with the cost of developing and administering a test that would place students more accurately.

Content validity, concurrent validity and predictive validity all have a part to play in the development of a test. For instance, in developing an English placement test for language schools, Hughes et al. (1996) validated test content against the content of three popular course books used by language schools in Britain, compared students' performance on the test with their performance on the existing placement tests of a number of language schools, and then examined the success of the test in placing students in classes. Only when this process was complete (and minor changes made on the basis of the results obtained) was the test published.

Other forms of evidence for construct validity

Investigations of a test's content validity and criterion-related validity provide evidence for its overall, or construct validity. However, they are not the only source of evidence. One could imagine a test that was meant to measure reading ability, the specifications for which included reference to a variety of reading sub-skills, including, for example, the ability to guess the meaning of

⁵ Because the full range of ability is not included, the validity coefficient is an underestimate (see previous footnote).

unknown words from the context in which they are met. Content validation of the test might confirm that these sub-skills were well represented in the test. Concurrent validation might reveal a strong relationship between students' performance on the test and their supervisors' assessment of their reading ability. But one would still not be sure that the items in the test were 'really' measuring the sub-skills listed in the specifications.

The word *construct* refers to any underlying ability (or trait) that is hypothesised in a theory of language ability. The ability to guess the meaning of unknown words from context, referred to above, would be an example. It is a matter of empirical research to establish whether or not such a distinct ability exists, can be measured, and is indeed measured in that test. Without confirming evidence from such research, it would not be possible to say that the part of a test that attempted to measure that ability has construct validity. If all of the items in a test were meant to measure specified abilities, then, without evidence that they were actually measuring those abilities, the construct validity of the whole test would be in question.

The reader may ask at this point whether such a demanding requirement for validity is appropriate for practical testing situations. It is easy to see the relevance of content validity in developing a test. And if a test has criterion-related validity, whether concurrent or predictive, surely it is doing its job well. But does it matter if we can't demonstrate that parts of the test are measuring exactly what we say they are measuring?

We have some sympathy for this view. What is more, we believe that gross, common-sense constructs like 'reading ability' and 'writing ability' are unproblematic. Similarly, the direct measurement of writing ability, for instance, should not cause us too much concern: even without research we can be fairly confident that we are measuring a distinct and meaningful ability (albeit a quite general and not closely defined ability)⁶. Once we try to measure such an ability indirectly, however, we can no longer take for granted what we are doing.

Let us imagine that we are indeed planning an indirect test of writing ability which must, for reasons of practicality, be multiple choice. We would need to begin by looking to a theory of writing ability for guidance as to the content and techniques that should be included in the test. This theory might tell us that underlying writing ability are a number of sub-abilities, such as control of punctuation, sensitivity to demands on style, and so on. We construct items that are meant to measure these sub-abilities and administer them as a pilot test. How do we know that this test really is measuring writing ability? One step we would almost certainly take is to obtain extensive samples of the writing ability of the group to

⁶ However, one may question the validity of the scales used to assess performance in, say, writing. How far do they reflect the development or acquisition of the skills they refer to? This may not be important in proficiency testing, where the scales may be based on levels of skill needed for a particular purpose (a job, for example). In achievement testing, scales that are not consistent with patterns of development may lack validity.

whom the test is first administered, and have these reliably scored. We would then compare scores on the pilot test with the scores given for the samples of writing. If there is a high level of agreement (and a coefficient of the kind described in the previous section can be calculated), then we have evidence that we are measuring writing ability with the test.

So far, however, although we may have developed a satisfactory indirect test of writing, we have not demonstrated the reality of the underlying constructs (control of punctuation, etc.). To do this we might administer a series of specially constructed tests, measuring each of the constructs by a number of different methods. In addition, compositions written by the people who took the tests could be scored separately for performance in relation to the hypothesised constructs (control of punctuation, for example). In this way, for each person, we would obtain a set of scores for each of the constructs. Coefficients could then be calculated between the various measures. If the coefficients between scores on the same construct are consistently higher than those between scores on different constructs, then we have evidence that we are indeed measuring separate and identifiable constructs. This knowledge would be particularly valuable if we wanted to use the test for diagnostic purposes.

Another way of obtaining evidence about the construct validity of a test is to investigate what test-takers actually do when they respond to an item. Two principal methods are used to gather such information: think aloud and retrospection. In the think aloud method, test-takers voice their thoughts as they respond to the item. In retrospection, they try to recollect what their thinking was as they responded. In both cases their thoughts are usually recorded, although a questionnaire may be used for the latter. The problem with the think aloud method is that the very voicing of thoughts may interfere with what would be the natural response to the item. The drawback to retrospection is that thoughts may be misremembered or forgotten. Despite these weaknesses, such research can give valuable insights into how items work (which may be quite different from what the test developer intended).

All test validation is to some degree a research activity. When it goes beyond content- and criterion-related validation, theories are put to the test and are confirmed, modified or abandoned. It is in this way that language testing can be put on a sounder, more scientific footing. But it will not all happen overnight; there is a long way to go. In the meantime, the practical language tester should try to keep abreast of what is known. When in doubt, where it is possible, direct testing of abilities is recommended.

Validity in scoring

It is worth pointing out that if a test is to have validity, not only the items but also the way in which the responses are scored must be valid. It is no use

having excellent items if they are scored invalidly. A reading test may call for short written responses. If the scoring of these responses takes into account spelling and grammar, then it is not valid (assuming the reading test is meant to measure just reading ability!). By measuring more than one ability, it makes the measurement of the one ability in question less accurate. There may be occasions when, because of misspelling or faulty grammar, it is not clear what the test-taker intended. In this case, the problem is with the item, not with the scoring. Similarly, if we are interested in measuring speaking or writing ability, it is not enough to elicit speech or writing in a valid fashion. The rating of that speech or writing has to be valid too. For instance, overemphasis on such mechanical features as spelling and punctuation can invalidate the scoring of written work (and so the test of writing).

Face validity

A test is said to have *face validity* if it *looks* as if it measures what it is supposed to measure. For example, a test that pretended to measure pronunciation ability but which did not require the test-taker to speak (and there have been some) might be thought to lack face validity. This would be true even if the test's construct and criterion-related validity could be demonstrated. Face validity is not a scientific notion and is not seen as providing evidence for construct validity, yet it can be very important. A test which does not have face validity may not be accepted by candidates, teachers, education authorities or employers. It may simply not be used; and if it is used, the candidates' reaction to it may mean that they do not perform on it in a way that truly reflects their ability. Novel techniques, particularly those which provide indirect measures, have to be introduced slowly, with care, and with convincing explanations.

How to make tests more valid

In the development of a high-stakes test, which may significantly affect the lives of those who take it, there is an obligation to carry out a full validation exercise before the test becomes operational.

In the case of teacher-made tests, full validation is unlikely to be possible. In these circumstances, we would recommend the following:

- First, write explicit specifications for the test (see Chapter 7) which take account of all that is known about the constructs that are to be measured. Make sure that you include a representative sample of the content of these in the test.
- Second, whenever feasible, use direct testing. If for some reason it is decided that indirect testing is necessary, reference should be made to the research literature to confirm that measurement of the relevant underlying constructs has been demonstrated using the testing techniques that are to be employed (this may often result in

disappointment, another reason for favouring direct testing!).

- Third, make sure that the scoring of responses relates directly to what is being tested.
- Finally, do everything possible to make the test reliable. If a test is not reliable, it cannot be valid. Reliability is dealt with in the next chapter.

Validity and fairness

It goes without saying that everyone wants language tests to be fair. Even the most hardnosed test professional would not deny the need for fairness. But what is fairness? And how do we achieve it?

The first essential for fairness is that a test be valid. Only if it measures accurately what it purports to measure, can it be fair. That is clear. But for fairness we need more. The test also has to be *used* fairly.

The fair use of tests has three components. First, all candidates have to be given an equal opportunity to show their ability on a test. This means that they are made familiar in advance with the structure of the test and the techniques used in it. They should also be provided with the opportunity to take a model version of the test and, if possible, be given feedback on their efforts. A handbook for the test (which among other things will provide sample items and scoring criteria) should be made available online or as hard copy (see Chapter 7).

Accommodation should be made in order not to disadvantage candidates with difficulties in hearing or speaking, with visual impairment, with specific learning difficulties such as dyslexia, or with other kinds of physical disability.

Second, the test and the scoring of the test should be conducted in appropriate conditions, with good-quality equipment where this is called for. See Chapter 18 for advice on test administration.

The third essential for fairness is that any test be used only for the purpose for which it is intended, and not for a purpose for which it was not designed. For example, the use of a test designed to measure language ability for academic purposes at university level should not be used as a general test of immigrants' language. This would be patently unfair, but it was happening in the United Kingdom at the time this was written.

Finally, test content should show sensitivity to all potential candidates' socio-cultural norms. To do otherwise might adversely affect candidates' performance and underestimate their ability.

Extended notions of validity

We have presented what we hope is a coherent, accessible and respectable account of validity. It has to be accepted, however, that there are language

testing theorists for whom the notion of validity goes beyond what we have described. For some of them at least, it is not the test itself, but the use to which it is actually put, that has (or does not have) validity, sometimes referred to as 'consequential validity'. The reader will recognise that this extended notion of validity is what in fact we have identified above as fairness⁷.

While we can all agree on the need for tests which measure accurately what they are intended to measure, and which are used in a defensible manner, our view is that it is not helpful to remove the possibility of discussing the validity of a test in itself, regardless of how it is actually used.

Last word

Test developers must make every effort to make their tests as valid as possible. Validation involves the collection of data of various kinds. Any published test should supply details of its validation, without which its validity (and suitability) can hardly be judged by a potential purchaser. Tests for which validity information is not available should be treated with caution.



READER ACTIVITIES

Consider any tests with which you are familiar. Assess each of them in terms of the various kinds of validity that have been presented in this chapter. What empirical evidence is there that the test is valid? If evidence is lacking, how would you set about gathering it?



FURTHER READING

The concept of validity

At first sight, validity seems a quite straightforward concept. On closer examination, however, it can seem impossibly complex, with some writers even finding it difficult to separate from the notion of reliability in some circumstances. In the present chapter, we have tried to present validity in a form which can be grasped by newcomers to the field and which will prove useful in thinking about and developing tests. For those who would like to explore the concept in greater depth, we would recommend: Anastasi and Urbina (1997) for a general discussion of test validity and ways of measuring it; Nitko (2001) for validity in the context of educational measurement; and Messick (1989) for a long, wide-ranging and detailed chapter on validity which is much cited in the language testing literature. His 1996 paper discusses the relationship between validity and backwash. Alderson et al. (1995) distinguish between internal and external categories of validity. Weir (2005) insists on a coherent validity framework as the basis of language test development. Extended notions of validity in language testing, and disagreements over them and their relation to fairness, are to be found in: Bachman and Palmer (2010), which presents a framework for the evaluation of assessment systems; Kane (2011) is a review of their

⁷ Except in the case where theorists include backwash as part of validity.

influential book. Also (in *Language Testing* 27, 2) Xi (2010), Davies (2010), Kane (2010) and Kunnan (2010).

Test validation

Fitzpatrick and Clenton (2010) throw light on a number of issues in test validation. A still interesting example of test validation (of the British Council *ELTS* test) in which a number of important issues are raised, is described and evaluated in Cripser and Davies (1988) and Hughes et al. (1988). Other accounts of validation can be found in Wall et al. (1994) and Fulcher (1997). Fox (2004) concerns the validation of an EAP test. Educational Testing Service (ETS) has conducted extensive research in the validation of its *TOEFL iBT®*, details of which can be found online. Alderson (2009) is a review of *TOEFL iBT®*. Pearson have published several validation reports online, including a research note by Riazi (2014), in which the correlation between *PTE Academic* total score and first semester GPA is reported as 0.34. Cumming and Berwick (1996) is a collection of papers on validation in language testing. For the argument (with which we do not agree) that there is no criterion against which 'communicative' language tests can be validated (in the sense of criterion-related validity), see Morrow (1986). Cohen (1984) describes early use of 'think-aloud' and retrospection. Buck (1991) and Wu (1998) provide examples of the use of introspection. Storey (1997) uses the think aloud technique. In a chapter on strategies used by test-takers, Cohen (2012) reports on more recent research into what candidates actually do when responding to test items. Bachman and Cohen (1998) is a collection of papers concerned with the relationship between second language acquisition and language testing research.

Content validity

Kim and Elder (2015) point to the importance of consulting domain specialists when constructing language tests for specific purposes. Weir et al. (1993) and Weir and Porter (1995) disagree with Alderson (1990a, 1990b) about the evidence for certain reading comprehension skills. Alderson and Kremmel (2013) warn against dependence on expert judgements, particularly when these involve categories which are themselves questionable. Stansfield and Hewitt (2005) discuss the effect of changing the pass score on the predictive validity of a test.

Face validity

Bradshaw (1990) investigates the face validity of a placement test.

Fairness

Taylor (2012) is a chapter on accommodation in language testing.

5

Reliability

Imagine that a hundred students take a 100-item test at three o'clock one Thursday afternoon. The test is not impossibly difficult or ridiculously easy for these students, so they do not all get zero or a perfect score of 100. Now what if, in fact, they had not taken the test on the Thursday but had taken it at three o'clock the previous afternoon? Would we expect each student to have got exactly the same score on the Wednesday as they actually did on the Thursday? The answer to this question must be no. Even if we assume that the test is excellent, that the conditions of administration are almost identical, that the scoring calls for no judgement on the part of the scorers and is carried out with perfect care, and that no learning or forgetting has taken place during the one-day interval, nevertheless we would not expect every individual to get precisely the same score on the Wednesday as they got on the Thursday. Human beings are not like that; they simply do not behave in exactly the same way on every occasion, even when the circumstances seem identical.

But if this is the case, it implies that we can never have complete trust in any set of test scores. We know that the scores would have been different if the test had been administered on the previous or the following day. This is inevitable, and we must accept it. What we have to do is construct, administer and score tests in such a way that the scores actually obtained on a test on a particular occasion are likely to be very similar to those which would have been obtained if it had been administered to the same students with the same ability, but at a different time. The more similar the scores would have been, the more *reliable* the test is said to be.

Look at the hypothetical data in Table 1(A). They represent the scores obtained by ten students who took a 100-item test (A) on a particular occasion, and those that they would have obtained if they had taken it a day later. Compare the two sets of scores. (Do not worry for the moment about the fact that we would never be able to obtain this information. Ways of estimating what scores people would have got on another occasion are discussed later. The most obvious of these is simply to have people take the same test twice.) Note the size of the difference between the two scores for each student.

TABLE 1(A): SCORES ON TEST A (INVENTED DATA)

Student	Score obtained	Score which would have been obtained on the following day
Bill	68	82
Mary	46	28
Ann	19	34
Harry	89	67
Cyril	43	63
Pauline	56	59
Don	43	35
Colin	27	23
Irene	76	62
Sue	62	49

Now look at Table 1(B), which displays the same kind of information for a second 100-item test (B). Again note the difference in scores for each student.

TABLE 1(B): SCORES ON TEST B (INVENTED DATA)

Student	Score obtained	Score which would have been obtained on the following day
Bill	65	69
Mary	48	52
Ann	23	21
Harry	85	90
Cyril	44	39
Pauline	56	59
Don	38	35
Colin	19	16
Irene	67	62
Sue	52	57

Which test seems the more reliable? The differences between the two sets of scores are much smaller for Test B than for Test A. On the evidence that we have here (and in practice we would not wish to make claims about reliability on the basis of such a small number of individuals), Test B appears to be more reliable than Test A.

Look now at Table 1(C), which represents scores of the same students on an interview using a five-point scale.

TABLE 1(C): SCORES ON INTERVIEW (INVENTED DATA)

Student	Score obtained	Score which would have been obtained on the following day
Bill	5	3
Mary	4	5
Ann	2	4
Harry	5	2
Cyril	2	4
Pauline	3	5
Don	3	1
Colin	1	2
Irene	4	5
Sue	3	1

In one sense the two sets of interview scores are very similar. The largest difference between a student's actual score and the one which would have been obtained on the following day is 3. But the largest *possible* difference is only 4! Really the two sets of scores are very different. This becomes apparent once we compare the size of the differences between students with the size of differences between scores for individual students. They are of about the same order of magnitude. The result of this can be seen if we place the students in order according to their interview score, the highest first.

TABLE 1(D): STUDENTS ORDERED ACCORDING TO SCORES

Actual score	Student	Score which would have been obtained on the following day	Student
5	Bill	5	Irene
	Harry		Mary
4	Mary	4	Pauline
	Irene		Ann
	Cyril		Cyril
3	Pauline	3	Bill
	Don		
	Sue		
2	Ann	2	Colin
	Cyril		Harry
1	Colin	1	Sue
			Don

The order based on their actual scores is markedly different from the one based on the scores they would have obtained if they had had the interview on the following day. This interview turns out in fact not to be very reliable at all.

The reliability coefficient

It is possible to quantify the reliability of a test in the form of a *reliability coefficient*. Reliability coefficients are like validity coefficients (Chapter 4).

They allow us to compare the reliability of different tests. The ideal reliability coefficient is 1. A test with a reliability coefficient of 1 is one which would give precisely the same results for a particular set of candidates regardless of when it happened to be administered. A test which had a reliability coefficient of zero (and let us hope that no such test exists!) would give sets of results quite unconnected with each other, in the sense that the score that someone actually got on a Wednesday would be no help at all in attempting to predict the score he or she would get if they took the test the day after. It is between the two extremes of 1 and zero that genuine test reliability coefficients are to be found.

Certain authors have suggested how high a reliability coefficient we should expect for different types of language tests. Lado (1961), for example, says that good vocabulary, structure and reading tests are usually in the 0.90 to 0.99 range, while auditory comprehension tests are more often in the 0.80 to 0.89 range. Oral production tests may be in the 0.70 to 0.79 range. He adds that a reliability coefficient of 0.85 might be considered high for a speaking production test but low for a reading test. These suggestions reflect what Lado sees as the different levels of difficulty the tester faces in achieving reliability in the testing of the different abilities, oral testing being the most difficult (see below and subsequent chapters for our views on this).

In fact the reliability coefficient that is to be sought will depend also on other considerations, most particularly the importance of the decisions that are to be taken on the basis of the test. The more high-stakes a test is, the greater reliability we must demand: for example, if we are to refuse someone the opportunity to study overseas because of their score on a language test, then we have to be pretty sure that their score would not have been much different if they had taken the test a day or two earlier or later. For a low-stakes test, such as a progress test, we can accept a lower level of reliability. The next section will explain how the reliability coefficient can be used to arrive at another figure (the standard error of measurement) to estimate likely differences of this kind. Before this is done, however, something has to be said about the way in which reliability coefficients are arrived at.

The first requirement is to have two sets of scores for comparison. The most obvious way of obtaining these is to get a group of subjects to take the same test twice. This is known as the *test-retest method*. The drawbacks are not difficult to see. If the second administration of the test is too soon after the first, then subjects are likely to recall items and their responses to them, making the same responses more likely and the reliability spuriously high. If there is too long a gap between administrations, then learning (or forgetting!) will have taken place, and the coefficient will be lower than it should be. However long the gap, the subjects are unlikely to be very motivated to take the same test twice, and this too is likely to have a depressing effect on the coefficient. These effects are reduced somewhat by the use of two different forms of the

same test (the *alternate forms method*). However, alternate forms are often simply not available.

It turns out, surprisingly, that the most common methods of obtaining the necessary two sets of scores involve only *one* administration of *one* test. Such methods provide us with a *coefficient of internal consistency*. The most basic of these is the *split half method*. In this the subjects take the test in the usual way, but each subject is given two scores. One score is for one half of the test, the second score is for the other half. The two sets of scores are then used to obtain the reliability coefficient as if the whole test had been taken twice. In order for this method to work, it is necessary for the test to be split into two halves which are really equivalent, through the careful matching of items (in fact where items in the test have been ordered in terms of difficulty, a split into odd-numbered items and even-numbered items may be adequate). It can be seen that this method is rather like the alternate forms method, except that the two 'forms' are only half the length¹.

It has been demonstrated empirically that this altogether more economical method will indeed give good estimates of alternate forms coefficients, provided that the alternate forms are closely equivalent to each other².

The standard error of measurement and the true score

While the reliability coefficient allows us to compare the reliability of tests, it does not tell us directly how close an individual's actual score is to what he or she might have scored on another occasion. With a little further calculation, however, it is possible to estimate how close a person's actual score is to what is called their *true score*. Imagine that it were possible for someone to take the same language test over and over again, an indefinitely large number of times, without their performance being affected by having already taken the test, and without their ability in the language changing. Unless the test is perfectly reliable, and provided that it is not so easy or difficult that the student always gets full marks or zero,

¹. Because of the reduced length, which will cause the coefficient to be less than it would be for the whole test, a statistical adjustment has to be made, using the Spearman-Brown formula (see Chapter 19).

². Note that a reliability coefficient can be misleading if there are even just a couple of candidates that score much higher (and/or much lower) than the others. The presence of such scores will cause the reliability coefficient to be misleadingly high. This is because the statistical methods used to estimate reliability compare the size of differences between candidates with the size of differences 'within' candidates (i.e. between candidates' two scores). The greater the relative difference between candidates, the greater will be the reliability coefficient. The difference between candidates will be exaggerated by the inclusion in the study of untypical candidates of the kind identified above. It is this which leads to an inappropriate estimate of reliability. See Nitko (2001) for details

we would expect their scores on the various administrations to vary. If we had all of these scores we would be able to calculate their average score, and it would seem not unreasonable to think of this average as the one that best represents the student's ability with respect to this particular test. It is this score, which for obvious reasons we can never know for certain, which is referred to as the candidate's true score.

We are able to make statements about the probability that a candidate's true score (the one which best represents their ability on the test) is within a certain number of points of the score they actually obtained on the test. In order to do this, we must first know the *standard error of measurement* of the particular test. The calculation of the standard error of measurement is based on the reliability coefficient and a measure of the spread of all the scores on the test (for a given spread of scores, the greater the reliability coefficient, the smaller will be the standard error of measurement). How such statements can be made using the standard error of measurement of the test is best illustrated by an example.

Suppose that a test has a standard error of measurement of 5. An individual scores 56₃ on that test. We are then in a position to make the following statements :

We can be about 68 percent certain that the person's true score lies in the range 51–61 (i.e. within one standard error of measurement of the score actually obtained on this occasion).

We can be about 95 percent certain that their true score is in the range 46–66 (i.e. within two standard errors of measurement of the score actually obtained).

We can be 99.7 percent certain that their true score is in the range 41–71 (i.e. within three standard errors of measurement of the score actually obtained).

These statements are based on what is known about the pattern of scores that would occur if it were in fact possible for someone to take the test repeatedly in the way described above. About 68 percent of their scores would be within one standard error of measurement, and so on. If in fact they only take the test once, we cannot be sure how their score on

³ These statistical statements are based on what is known about the way a person's scores would tend to be distributed if they took the same test an indefinitely large number of times (without the experience of any test-taking occasion affecting performance on any other occasion). The scores would follow what is called a normal distribution (see Woods et al. 1986, for discussion beyond the scope of the present book). It is the known structure of the normal distribution which allows us to say what percentage of scores will fall within a certain range (for example about 68 percent of scores will fall within one standard error of measurement of the true score). Since about 68 percent of actual scores will be within one standard error of measurement of the true score, we can be about 68 percent certain that any particular actual score will be within one standard error of measurement of the true score.

that occasion relates to their true score, but we are still able to make probabilistic statements as above⁴.

In the end, the statistical rationale is not important. What is important is to recognise how we can use the standard error of measurement to inform decisions that we take on the basis of test scores. We should, for example, be very wary of taking important negative decisions about people's future if the standard error of measurement indicates that their true score is quite possibly equal to or above the score that would lead to a positive decision, even though their actual score is below it. For example, someone needs a score of 90 in order to study at an English-medium university but only scores 88 on the test. Let us say that the test has a reported standard error of measurement of 4.5. This means that there is a 68 percent chance that the person's true score is somewhere between 83.5 and 92.5. In these circumstances, it would be unwise to automatically deny the person entry to the university. Where possible, other information about the candidate should be sought and taken into account before making a decision⁵.

In order to help informed decisions to be made, all published tests should provide users with not only the reliability coefficient but also the standard error of measurement.

A more recent approach to the statistical analysis of test data, known as *Item Response Theory (IRT)*, allows an even better estimate of how far an individual test-taker's actual score is likely to diverge from their true score. While classical analysis gives us a single estimate for all test-takers, IRT gives an estimate for each individual, basing this estimate on that individual's performance on each of the items on the test. Examples of this estimate, usually referred to as the *standard error* of the individual's score, can be found in Chapter 19.

IRT is particularly useful, some might say essential, in computer adaptive testing (Chapter 3). Using IRT, after each item has been responded to by an individual, an estimate is made of the standard error, and this is repeated with each successive item until what has previously been set as the required standard error is reached. At that point, testing ends and the individual's score is recorded.

What has been said so far in this chapter has concerned itself with the consistency of *scores* that candidates obtain on a test. In criterion-referenced testing, we are often less interested in scores than in whether

⁴. It should be clear that there is no such thing as a 'good' or a 'bad' standard error of measurement. It is the particular use made of particular scores in relation to a particular standard error of measurement which may be considered acceptable or unacceptable.

⁵. As indicated in the previous chapter, there is a growing movement towards taking multiple measures of ability. In our view, these are most important in high-stakes tests and at the pass/fail margin. Non-testing information of the kind described in Chapter 16 can make an important contribution to decision making, provided that the possible limits of its reliability are taken into account.

a candidate has reached the criterion which has been set. In this case, the consistency which we are looking for is referred to as *decision consistency* (rather than reliability)⁶.

We want to know whether a test is consistent in deciding whether or not the candidates have or have not reached the criterion. Imagine a case where 50 candidates take a test (perhaps two alternate forms of it) twice. Those who reach a criterion may be called *masters* (in the sense of having mastered the skills, or whatever, that are being tested) and those who do not reach it may be called *non-masters*. Of the 50 candidates:

18 are masters on both occasions

15 are non-masters on both occasions

9 are masters on the first occasion but non-masters on the second

8 are non-masters on the first occasion but masters on the second

So, out of 50 candidates, 33 are assigned to the same category (master or non-master on both occasions). Thirty-three out of 50 can be expressed as a percentage (66%) or as a proportion (0.66). This last value, 0.66, is known as the *percent agreement* and is an accepted estimate of decision consistency. For other methods for estimating decision consistency (and they are not limited to just two groups, masters and non-masters), see the Further reading section.

We have seen the importance of reliability. If a test is not reliable, then we know that the actual scores of many individuals are likely to be quite different from their true scores. This means that we can place little reliance on those scores. Even where reliability is quite high, the standard error of measurement (or the standard errors obtained through IRT) serves to remind us that in the case of some individuals there is quite possibly a large discrepancy between actual score and true score. This should make us very cautious about making important decisions on the basis of the test scores of candidates whose actual scores place them close to the cut-off point (the point that divides 'passes' from 'fails'). We should at least consider the possibility of gathering further relevant information on the language ability of such candidates.

Having seen the importance of reliability, we shall consider, later in the chapter, how to make our tests more reliable. Before that, however, we shall look at another aspect of reliability.

Scorer reliability

In the first example given in this chapter we spoke about scores on a multiple choice test. It was most unlikely, we thought, that every candidate

⁶ A criterion-referenced test may be very consistent yet yield a low reliability coefficient. This is because candidates' scores, although they classify the candidates consistently, may be very limited in range (see footnote 2). For this reason, it is recommended that one should use methods specifically designed for criterion-referenced tests.

would get precisely the same score on both of two possible administrations of the test. We assumed, however, that scoring of the test would be 'perfect'. That is, if a particular candidate did perform in exactly the same way on the two occasions, they would be given the same score on both occasions. That is, any one scorer would give the same score on the two occasions, and this would be the same score as would be given by any other scorer on either occasion⁷.

It is possible to quantify the level of agreement given by the same or different scorers on different occasions by means of a *scorer reliability coefficient* which can be interpreted in a similar way to the test reliability coefficient. In the case of the multiple choice test just described, the scorer reliability coefficient would be 1. As we noted in Chapter 3, when scoring requires no judgement, and could in principle or in practice be carried out by a computer, the test is said to be objective. Only carelessness should cause the scorer reliability coefficients of objective tests to fall below 1.

However, we did not make the assumption of perfectly consistent scoring in the case of the interview scores discussed earlier in the chapter. It would probably have seemed to the reader an unreasonable assumption. We can accept that scorers should be able to be consistent when there is only one easily recognised correct response. But when a degree of judgement is called for on the part of the scorer, as in the scoring of performance in an interview, perfect consistency is not to be expected. Such subjective tests will not have scorer reliability coefficients of 1! Indeed there was a time when many people thought that scorer reliability coefficients (and also the reliability of the test) would always be too low to justify the use of subjective measures of language ability in serious language testing. This view is less widely held today. While the perfect reliability of objective tests is not obtainable in subjective tests, there are ways of making it sufficiently high for test results to be valuable. It is possible, for instance, to obtain scorer reliability coefficients of over 0.9 for the scoring of written compositions.

It is perhaps worth making explicit something about the relationship between scorer reliability and test reliability. If the scoring of a test is not reliable, then the test results cannot be reliable either. Indeed the test reliability coefficient will almost certainly be lower than scorer reliability, since other sources of unreliability will be additional to what enters through imperfect scoring. In a case we know of, the scorer reliability coefficient on a composition writing test was 0.92, while the reliability coefficient for the test was 0.84. Variability in the performance of individual candidates accounted for the difference between the two coefficients.

⁷ The reliability of one person scoring the same test responses on different occasions is called 'intra-scorer reliability'; the reliability of different people scoring the same test responses is called 'inter-scorer reliability'.

How to make tests more reliable

As we have seen, there are two components of test reliability: the performance of candidates from occasion to occasion, and the reliability of the scoring. We will begin by suggesting ways of achieving consistent performances from candidates and then turn our attention to improving scorer reliability.

Take enough samples of behaviour

Other things being equal, the more items that you have on a test, the more reliable that test will be. This seems intuitively right. If we wanted to know how good an archer someone was, we wouldn't rely on the evidence of a single shot at the target. That one shot could be quite unrepresentative of their ability. To be satisfied that we had a really reliable measure of the ability we would want to see a large number of shots at the target.

The same is true for language testing. It has been demonstrated empirically that the addition of further items will make a test more reliable. There is even a formula (the Spearman–Brown formula, see Chapter 19) that allows one to estimate how many extra items similar to the ones already in the test will be needed to increase the reliability coefficient to a required level. One thing to bear in mind, however, is that the additional items should be independent of each other and of existing items. Imagine a reading test that asks the question: 'Where did the thief hide the jewels?' If an additional item following that took the form, 'What was unusual about the hiding place?', it would not make a full contribution to an increase in the reliability of the test. Why not? Because it is hardly possible for a candidate who got the original question wrong to get the supplementary question right. Such a candidate is effectively prevented from answering the additional question; for that candidate, in reality, there is no additional question. We do not get an additional sample of their behaviour, so the reliability of our estimate of their ability is not increased.

Each additional item should as far as possible represent a fresh start for the candidate. By doing this we are able to gain additional information on all of the candidates – information that will make test results more reliable. The use of the word 'item' should not be taken to mean only brief questions and answers. In a test of writing, for example, where candidates have to produce a number of passages, each of those passages is to be regarded as an item. The more independent passages there are, the more reliable will be the test. In the same way, in an interview used to test oral ability, the candidate should be given as many 'fresh starts' as possible. More detailed implications of the need to obtain sufficiently large samples of behaviour will be outlined later in the book, in chapters devoted to the testing of particular abilities.

While it is important to make a test long enough to achieve satisfactory reliability, it should not be made so long that the candidates become so bored or tired that the behaviour they exhibit becomes unrepresentative of their ability. At the same time, it may often be necessary to resist pressure to make a test shorter than is appropriate. The usual argument for shortening a test is that it is not practical for it to be longer. The answer to this is that accurate information does not come cheaply: if such information is needed, then the price has to be paid. In general, the more important the decisions based on a test, the longer the test should be. Jephthah used the pronunciation of the word 'shibboleth' as a test to distinguish his own men from Ephraimites, who could not pronounce *sh*. Those who failed the test were executed. Any of Jephthah's own men killed in error might have wished for a longer, more reliable test.

Exclude items which do not discriminate well between weaker and stronger students

Items on which strong students and weak students perform with similar degrees of success contribute little to the reliability of a test. Statistical analysis of items (Chapter 19) will reveal which items do not discriminate well. These are likely to include items which are too easy or too difficult for the candidates, but not only these. Normally, such items should be removed from the test and replaced with items which discriminate better. That said, a small number of easy, non-discriminating items may be kept at the beginning of a test to give candidates confidence and reduce the stress they feel.

Do not allow candidates too much freedom

In some kinds of language test there is a tendency to offer candidates a choice of questions and then to allow them a great deal of freedom in the way that they answer the ones that they have chosen. An example would be a test of writing where the candidates are simply given a selection of titles from which to choose. Such a procedure is likely to have a depressing effect on the reliability of the test. The more freedom that is given, the greater is likely to be the difference between the performance actually elicited and the performance that would have been elicited had the test been taken, say, a day later. In general, therefore, candidates should not be given a choice, and the range over which possible answers might vary should be restricted. Compare the following writing tasks:

1. Write a composition on tourism.
2. Write a composition on tourism in this country.
3. Write a composition on how we might develop the tourist industry in this country.

4. Discuss the following measures intended to increase the number of foreign tourists coming to this country:
 - i) More/Better advertising and/or information (Where? What form should it take?).
 - ii) Improve facilities (hotels, transportation, communication, etc.).
 - iii) Training of personnel (guides, hotel managers, etc.).

The successive tasks impose more and more control over what is written. The fourth task is likely to be a much more reliable indicator of writing ability than the first. The general principle of restricting the freedom of candidates will be taken up again in chapters relating to particular skills. It should perhaps be said here, however, that in restricting the students we must be careful not to distort too much the task that we really want to see them perform. The potential tension between reliability and validity is addressed at the end of the chapter.

Write unambiguous items

It is essential that candidates should not be presented with items whose meaning is not clear or to which there is an acceptable answer which the test writer has not anticipated. In a reading test we once set the following open-ended question, based on a lengthy reading passage about English accents and dialects: Where does the author direct the reader who is interested in non-standard dialects of English? The expected answer was the Further reading section of the book. A number of candidates answered 'page 3', which was the place in the text where the author actually said that the interested reader should look in the Further reading section. Only the alertness of those scoring the test revealed that there was a completely unanticipated correct answer to the question. If that had not happened, a correct answer would have been scored as incorrect. The fact that an individual candidate might interpret the question in different ways on different occasions means that the item is not contributing fully to the reliability of the test.

The best way to arrive at unambiguous items is, having drafted them, to subject them to the critical scrutiny of colleagues, who should try as hard as they can to find alternative interpretations to the ones intended. If this task is entered into in the right spirit – one of good-natured collegiality – most of the problems can be identified before the test is administered. Pre-testing of the items on a group of people comparable to those for whom the test is intended (see Chapter 7) should reveal the remaining problems. Where pre-testing is not practicable, scorers must be on the lookout for patterns of response that indicate that there are problem items.

Provide clear and explicit instructions

This applies both to written and oral instructions. If it is possible for candidates to misinterpret what they are asked to do, then on some occasions some of them certainly will. It is by no means always the

weakest candidates who are misled by ambiguous instructions; indeed it is often the better candidate who is able to provide the alternative interpretation. A common fault of tests written for the students of a particular teaching institution is the supposition that the students all know what is intended by carelessly worded instructions. The frequency of the complaint that students are unintelligent, have been stupid, or have wilfully misunderstood what they were asked to do, reveals that the supposition is often unwarranted. Test writers should not rely on the students' powers of telepathy to elicit the desired behaviour. Again, the use of colleagues to criticise drafts of instructions (including those which will be spoken) is the best means of avoiding problems. Spoken instructions should always be read from a prepared script in order to avoid introducing confusion.

Ensure that tests are well laid out and perfectly legible

Too often, institutional tests are badly typed (or handwritten), have too much text in too small a space, and are poorly reproduced. As a result, students are faced with additional tasks which are not ones meant to measure their language ability. Their variable performance on the unwanted tasks will lower the reliability of a test.

Make candidates familiar with format and testing techniques

If any aspect of a test is unfamiliar to candidates, they are likely to perform less well than they would do otherwise (on subsequently taking a parallel version, for example). For this reason, every effort must be made to ensure that all candidates have the opportunity to learn just what will be required of them. This may mean the distribution of sample tests (or of past test papers), or at least the provision of practice materials in the case of tests set within teaching institutions.

Provide uniform and non-distracting conditions of administration

The greater the differences between one administration of a test and another, the greater the differences one can expect between a candidate's performance on the two occasions. Great care should be taken to ensure uniformity. For example, timing should be specified and strictly adhered to; the acoustic conditions should be similar for all administrations of a listening test. Every precaution should be taken to maintain a quiet setting with no distracting sounds or movements.

We turn now to ways of obtaining scorer reliability, which is essential to test reliability.

Use items that permit scoring which is as objective as possible

This may appear to be a recommendation to use multiple choice items, which permit completely objective scoring. This is not intended. While it

would be a mistake to say that multiple choice items are never appropriate, it is certainly true that there are many circumstances in which they are quite inappropriate. What is more, good multiple choice items are notoriously difficult to write and always require extensive pre-testing. A substantial part of Chapter 8 is given over to the shortcomings of the multiple choice technique.

An alternative to multiple choice is the open-ended item which has a unique, possibly one-word, correct response which the candidates produce themselves. This too should ensure objective scoring, but in fact problems with such matters as spelling which makes a candidate's meaning unclear (say, in a listening test) often make demands on the scorer's judgement. The longer the required response, the greater the difficulties of this kind. One way of dealing with this is to structure the candidate's response by providing part of it. For example, the open-ended question, *What was different about the results?* may be designed to elicit the response, *Success was closely associated with high motivation.* This is likely to cause problems for scoring. Greater scorer reliability will probably be achieved if the question is followed by:

..... was closely associated with

Items of this kind are discussed in later chapters.

Make comparisons between candidates as direct as possible

This reinforces the suggestion already made that candidates should not be given a choice of items and that they should be limited in the way that they are allowed to respond. Scoring the compositions all on one topic will be more reliable than if the candidates are allowed to choose from six topics, as has been the case in some well-known tests. The scoring should be all the more reliable if the compositions are guided as in the example above, in the section, 'Do not allow candidates too much freedom'.

Provide a detailed scoring key

This should specify acceptable answers and assign points for acceptable partially correct responses. For high scorer reliability the key should be as detailed as possible in its assignment of points. It should be the outcome of efforts to anticipate all possible responses and have been subjected to group criticism. (This advice applies only where responses can be classed as partially or totally 'correct', not in the case of compositions, for instance.)

Train scorers

This is especially important where scoring is most subjective. The scoring of compositions, for example, should not be assigned to anyone who has

not learned to score accurately compositions from past administrations. After each administration, patterns of scoring should be analysed. Individuals whose scoring deviates markedly and inconsistently from the norm should not be used again.

Agree acceptable responses and appropriate scores at outset of scoring

A sample of scripts should be taken immediately after the administration of the test. Where there are compositions, archetypal representatives of different levels of ability should be selected. Only when all scorers are agreed on the scores to be given to these should real scoring begin.

Having said that, we should add that for the scoring of compositions an alternative approach, known as *comparative judgement*, has gained currency in recent years. From the outset, each judge is presented with a pair of scripts on screen and asked simply to say which is the better of the two. This process is repeated over and over, with multiple judges, and the comparative judgement algorithm combines all the decisions and uses them to create a measurement scale, so all the scripts can be placed on this single scale. It is reported that the method results in high reliability. More will be said in Chapter 9 about the scoring of compositions.

For short-answer questions, the scorers should note any difficulties they have in assigning points (the key is unlikely to have anticipated every relevant response), and bring these to the attention of whoever is supervising that part of the scoring. Once a decision has been taken as to the points to be assigned, the supervisor should convey it to all the scorers concerned.

Identify candidates by number, not name

Scorers inevitably have expectations of candidates that they know. Except in purely objective testing, this will affect the way that they score. Studies have shown that even where the candidates are unknown to the scorers, the name on a script (or a photograph) will make a significant difference to the scores given. For example, a scorer may be influenced by the gender or nationality of a name into making predictions which can affect the score given. The identification of candidates only by number will reduce such effects.

Employ multiple, independent scoring

As a general rule, and certainly where testing is subjective, all scripts should be scored by at least two independent scorers. Neither scorer

should know how the other has scored a test paper. Scores should be recorded on separate score sheets and passed to a third, senior, colleague, who compares the two sets of scores and investigates discrepancies.

Reliability and validity

To be valid a test must provide consistently accurate measurements. It must therefore be reliable. A reliable test, however, may not be valid at all. For example, as a writing test we could require candidates to write down the translation equivalents of 500 words in their own language. This might well be a reliable test; but it is unlikely to be a valid test of writing.

In our efforts to make tests reliable, we must be wary of reducing their validity, as happens when multiple choice items are used inappropriately. Earlier in this chapter it was admitted that restricting the scope of what candidates are permitted to write in a composition might diminish the validity of the task. This depends in part on what exactly we are trying to measure by setting the task. If we are interested in candidates' ability to structure a composition, then it would be hard to justify providing them with a structure in order to increase reliability. At the same time we would still try to restrict candidates in ways which would not render their performance on the task invalid.

There will always be some tension between reliability and validity. The tester has to balance potential gains in one against losses in the other.



READER ACTIVITIES

1. What published tests are you familiar with? Try to find out their reliability coefficients. What method was used to arrive at these? What are the standard errors of measurement?
2. The *TOEFL*® internet-based test is reported as having a standard error of measurement of 4.26 on a typical administration. A particular American college states that it requires a score of 100 on the test for entry. What would you think of students applying to that college and making scores of 104, 100, 96, or 90?
3. Look at your own institutional tests. Using the list of points in the chapter, say in what ways you could improve their reliability.
4. What examples can you think of where there would be a tension between reliability and validity? In cases that you know, do you think the right balance has been struck?



FURTHER READING

For more on reliability in general and the relationship between different estimates of reliability and the different factors that account for it, see Anastasi and Urbina (1997). For reliability in educational measurement see Nitko (2001) and Feldt and Brennan's chapter in Linn (1989). The latter explains the application of generalisability theory, which lets us calculate the relative contributions of different sources of unreliability (e.g. different versions of a test, different scorers, etc.). We should, however, warn less mathematically minded readers that their chapter is highly technical.

For four 'relatively easy to calculate' estimates of decision consistency see Brown (1990). For further discussion of consistency in criterion-referenced testing, see Brown and Hudson (2002) and Nitko (2001). For what we think is an exaggerated view of the difficulty of achieving high reliability in more communicative tasks, see Lado (1961). This may have been written more than fifty years ago, but the same beliefs are still expressed today.

6 Achieving positive backwash

Backwash is the effect that tests have on learning and teaching. Before the first edition of this book appeared, little attention was given to the subject. By the time of the second edition, there was much more interest in the topic. Backwash was established as an important part of the impact that a test may have on learners and teachers, on educational systems, and on society at large. Calls had been made for explicit models of backwash, and research had begun into the processes by which it might be achieved¹.

Now, we are happy to say, we can read the results of research that has confirmed and quantified the effect of tests on teaching and learning. The Further reading section provides a guide to that research. We have also been encouraged by seeing the efforts of major language testing institutions (such as ETS in the United States and Cambridge Assessment English in the UK) to change their tests in ways that will encourage positive backwash.

We have no doubt that over the next few years continuing research into backwash will result in a better understanding of the processes involved and how different variables contribute to its effect in different situations. Nevertheless, we believe that the advice which follows, based largely on our practical experience, will prove helpful to teachers seeking to create positive backwash in their own situation.

Test the abilities whose development you want to encourage

For example, if you want to encourage oral ability, then test oral ability². This is very obvious, yet it is surprising how often it has not been done. There is a tendency to test what is easiest to test rather than what is most important to test. Reasons for not testing particular abilities may take many forms. It is often said, for instance, that sufficiently high reliability cannot be obtained when a form of testing (such as an oral interview) requires subjective scoring. This is simply not the case, and in addition to the advice already given in the previous chapter, more detailed suggestions for achieving satisfactory reliability of subjective tests are to be found in

¹ The word 'washback' is being increasingly used in place of 'backwash'. We will continue to use the original term 'backwash', except when citing other authors.

² Bearing in mind what was said in Chapter 4, it is important that the scoring or rating of test performance (as well as the means of elicitation) should be valid.

Chapters 9 and 10. The other most frequent reason given for not testing is the expense involved in terms of time and money. This is discussed later in the chapter.

It is important not only that certain abilities should be tested, but also that they should be given sufficient weight in relation to other abilities. One of us well remembers his French teacher telling the class that, since the oral component of the General Certificate of Education examination in French (which we were to take later in the year) carried so few marks, we should not waste our time preparing for it. The examining board concerned was hardly encouraging positive backwash.

Sample widely and unpredictably

Normally a test can measure only a sample of everything included in the specifications. It is important that this sample should represent as far as possible the full scope of what is specified. If not, if the sample is taken from only a restricted area of the specifications, then the backwash effect will tend to be felt only in that area. If, for example, the specifications for a writing test include three or more kinds of task, but repeatedly, over the years, versions of the test include only the same two kinds of task (for instance: compare/contrast; describe/interpret a chart or graph), the likely outcome is that much preparation for the test will be limited to those two types of task. The backwash effect may not be as positive as it might have been had a wider range of tasks been used.

Whenever the content of a test becomes highly predictable, teaching and learning are likely to concentrate on what can be predicted. An effort should therefore be made to test across the full range of the specifications (in the case of achievement tests, this should be equivalent to a fully elaborated set of objectives), even where this involves elements that lend themselves less readily to testing³.

We must add that core elements of the specifications (those which we believe are most important) should always be represented in each version of a test.

Use direct testing

As we saw in Chapter 3, direct testing implies the testing of performance skills, with texts and tasks as authentic as possible. If we test directly the skills that we are interested in fostering, then practice for the test

³ It has to be admitted that high-stakes tests will always attract entrepreneurs who offer training courses that attempt to provide potential candidates with tricks and forms of words that will enable them to make higher scores, without necessarily improving their language abilities. This kind of training hardly represents positive backwash. The aim of test constructors must be to minimise the possibility of such training being successful.

will naturally involve practice in those skills. If we want people to learn to write compositions, we should get them to write compositions in the test. If a course objective is that students should be able to read scientific articles, then we should get them to do that in the test. Immediately we begin to test indirectly, we are removing an incentive for students to practise in the way that we want them to.

Make testing criterion-referenced

If test specifications make clear what candidates have to be able to do, and with what degree of success, then students will have a clear picture of what they have to achieve. What is more, they will know that if they do perform the tasks at the criterial level, then they will be successful on the test, regardless of how other students perform. Both these things will help to motivate students. Where testing is not criterion-referenced, it becomes easy for teachers and students to assume that a certain (perhaps very high) percentage of candidates will pass, almost regardless of the absolute standard that they reach.

The possibility exists of having a series of criterion-referenced tests, each representing a different level of achievement or proficiency. The tests are constructed such that a 'pass' is obtained only by completing the great majority of the test tasks successfully. Students are required to take only the test (or tests) on which they are expected to be successful. As a result, they are spared the dispiriting, demotivating experience of taking a test on which they can, for example, respond correctly to fewer than half of the items (and yet be given a pass). This type of testing, we believe, should encourage positive attitudes to language learning. At one time it was the basis of some GCSE (General Certificate of Secondary Education) examinations in Britain.

It has to be admitted that there is one potential drawback to having a series of criterion-referenced tests for which a candidate is entered for only one of them. Someone has to decide which test to take. Whether it is the candidate, a teacher, or some other adviser, mistakes may be made. The candidate's ability may be underestimated or overestimated, resulting in the candidate taking an inappropriate test. One solution to this problem would be to have a single computer adaptive test. This could work well for a test of grammar or vocabulary. For a test of writing, however, where extended pieces of writing are called for, it is hard to see how that would work, unless initial items were short in nature and computer-scoreable. These initial items would effectively form a brief screening test and would serve to direct candidates to longer items. Traditional tests of speaking, carried out with a human interlocutor, are, or should be, adaptive in nature.

Base achievement tests on objectives

If achievement tests are based on objectives, rather than on detailed teaching and textbook content, they will provide a truer picture of what

has actually been achieved. Teaching and learning will tend to be evaluated against those objectives. As a result, there will be constant pressure to achieve them. This was argued more fully in Chapter 3.

Ensure the test is known and understood by students and teachers

However good the potential backwash effect of a test may be, the effect will not be fully realised if students and teachers do not know and understand what the test demands of them. The rationale for the test, its specifications, and sample items (including examples of written and oral performance with grades and examiner comments) should be made available to everyone concerned with preparation for the test. This is particularly important when a new test is being introduced, especially if it incorporates novel testing methods. Another, equally important, reason for supplying information of this kind is to increase test reliability, as was noted in the previous chapter.

Where necessary, provide assistance to teachers

The introduction of a new test may make demands on teachers to which they are not equal. If, for example, a longstanding national test of grammatical structure and vocabulary is to be replaced by a direct test of a much more communicative nature, it is possible that many teachers will feel that they do not know how to teach communicative skills. One important reason for introducing the new test may have been to encourage communicative language teaching, but if the teachers need guidance and possibly training, and these are not given, the test will not achieve its intended effect. It may simply cause chaos and disaffection. Where new tests are meant to help change teaching, support has to be given to help effect the change.

Counting the cost

One of the desirable qualities of tests which trips quite readily off the tongue of many testers, after validity and reliability, is that of practicality. Other things being equal, it is good that a test should be easy and cheap to construct, administer, score and interpret. We should not forget that testing costs time and money that could be put to alternative uses.

It is unlikely to have escaped the reader's notice that at least some of the recommendations listed above for creating positive backwash involve more than minimal expense. The individual direct testing of some abilities will take a great deal of time, as will the reliable scoring of performance on any subjective test. The production and distribution of sample tests and the training of teachers will also be costly. It might be argued, therefore,

that such procedures are impractical. In our opinion, this would reveal an incomplete understanding of what is involved. Before we decide that we cannot afford to test in a way that will promote positive backwash, we have to ask ourselves a question: What will be the cost of *not* achieving positive backwash? When we compare the cost of the test with the waste of effort and time on the part of teachers and students in activities quite inappropriate to their true learning goals (and in some circumstances, with the potential loss to the national economy of not having more people competent in foreign languages), we are likely to decide that we cannot afford not to introduce a test with a powerful positive backwash effect.



READER ACTIVITIES

1. How would you improve the backwash effect of tests that you know? Be as specific as possible. (This is a follow-up to Activity 1 at the end of Chapter 1.)
2. Rehearse the arguments you would use to convince a sceptic that it would be worthwhile making the changes that you recommend.



FURTHER READING

Theoretical issues

Alderson and Wall (1993) question the existence of backwash.

Language Testing 13, 3 (1996) is a special issue devoted to backwash. In it Messick discusses backwash in relation to validity. Bailey (1996) reviews the concept of backwash in language testing, including Hughes's (1993) proposed model and Alderson and Wall's (1993) fifteen hypotheses about backwash. Wall (1996) looks to developments in general education and to innovation theory for insights into backwash.

Hamp-Lyons's (1997a) article raises ethical concerns in relation to backwash, impact and validity. Her 1997b article discusses ethical issues in test preparation practice for *TOEFL*®, to which Wadden and Hilke (1999) take exception. Hamp-Lyons (1999) responds to their criticisms.

Brown and Hudson (1998) lay out the assessment possibilities for language teachers and argue that one of the criteria for choice of assessment method is potential backwash effect. Alderson (2009) reviews the new *TOEFL*® and comments on its potential for positive backwash.

Research into backwash

Wall and Alderson (1993) investigate backwash in a project in Sri Lanka with which they were concerned, argue that the processes involved in backwash are not straightforward, and call for a model of backwash and for further research. Shohamy et al. (1996) report that two different tests have different patterns of backwash. Watanabe (1996) investigates the possible effect of university entrance examinations in Japan on classroom methodology. Alderson and Hamp-Lyons (1996) report on a study into *TOEFL*® preparation courses and backwash. Muñoz and Álvarez (2010) is an account of a successful attempt to create positive backwash in a

Colombian university. Cheng (2005) reports on her research into backwash in Hong Kong. Cheng et al. (2011) report on the impact of introducing teachers' assessments as part of a high-stakes exam. Choi (2008) reports on the negative backwash effects of standardised multiple choice tests in the Korean education system. Luxia (2005) examines the failure of a high-stakes test to achieve its intended backwash effects. Saif (2006) describes an attempt to achieve positive backwash. Cheng et al. (2004) is a collection of articles on carrying out research into backwash. Cheng and Curtis (2012) summarise the results of research into backwash and make recommendations for future research. Green (2007) reports on research into the effect of the academic writing module of a major test on preparation for university study (*IELTS*). Wall and Horák (2006, 2008, 2011) is a series of reports on the impact of the new *TOEFL*® on teaching and learning. All of their reports are available online.

7

Stages of test development

This chapter presents a set of procedures for the construction of a language test within a teaching institution or organisation. Subsequent chapters deal with the testing of individual language skills and components of language. We begin by outlining the procedures, before describing their implementation in the development of an achievement test and a placement test.

The procedures we recommend are listed below.



PROCEDURES IN TEST DEVELOPMENT

1. Make a full and clear statement of the testing 'problem'.
2. Draft a complete set of specifications for the test.
3. Submit draft specifications to experts and stakeholders for feedback.
4. Revise specifications.
5. On the basis of the revised specifications, write and moderate items.
6. Trial the items informally on expert speakers¹ and reject or modify problematic ones as necessary.
7. Trial the test on a group of non-expert speakers similar to those for whom the test is intended.
8. Analyse the results of the trial and make any necessary changes.
9. Calibrate scales.
10. Carry out validation.
11. Write handbooks for test-takers, test users and staff.

Before looking more closely at this set of procedures, it is worth saying that test development is best carried out by a team. It is very difficult for a single individual to develop a successful test, if only because of the need to look objectively at what is being proposed at each stage of development. This difficulty can be seen most clearly at the stage of item writing, when faults in an item which are obvious to others are often invisible to the person who wrote the item. Writing items is a creative process, and we tend to think of our items as minor works of art or even, it sometimes seems, our babies. We do not find it easy to admit that our baby is not as beautiful as we had thought. One of the qualities to be looked for in item writers, therefore, is a willingness to accept justified criticism of the items which they have written. Other desirable qualities – not only for item writers but for test developers in general – are: expert command of the

¹ Because of the widespread rejection of the notion of a 'native speaker' we will use 'expert speaker' to refer to someone who is completely proficient in a language.

language, intelligence and imagination (to create contexts in items and to foresee possible misinterpretations).

1. Stating the problem

It cannot be said too many times that the essential first step in testing is to make oneself perfectly clear about what it is one wants to know and for what purpose. The following questions, the significance of which should be clear from previous chapters, have to be answered:

- i. What kind of test is it to be? Achievement (final or progress), proficiency, diagnostic or placement?
- ii. What is its precise purpose?
- iii. What abilities are to be tested?
- iv. How detailed must the results be?
- v. How accurate must the results be?
- vi. How important is backwash?
- vii. What constraints are set by unavailability of expertise, facilities, time (for construction, administration and scoring)?

Once the problem is clear, steps can be taken to solve it. It is to be hoped that a handbook of the present kind will take readers a long way towards appropriate solutions. In addition, however, efforts should be made to gather information on tests that have been designed for similar situations. If possible, samples of such tests should be obtained. There is nothing dishonourable in doing this; it is what professional testing bodies do when they are planning a test of a kind for which they do not already have first-hand experience. Nor does it contradict the claim made earlier that each testing situation is unique. It is not intended that other tests should simply be copied; rather that their development can serve to suggest possibilities and to help avoid the need to 'reinvent the wheel'.

2. Writing specifications for the test

A set of specifications for the test must be written at the outset². This will include information on: content, test structure, timing, medium/channel, techniques to be used, criterial levels of performance, and scoring procedures.

². This does not mean that the specifications should never be modified. Trialling may reveal, for example, that there are too many items to be responded to in the time assigned to them. The circumstances in which the test is to be administered may change. It is also true that at the time of writing specifications certain details may be unknowable. For example, we may not know how many items will be needed in a test in order to make it reliable and valid for its purpose.

i. Content

This refers not to the content of a single, particular version of a test, but to the entire potential content of any number of versions. Samples of this content will appear in individual versions of the test.

The fuller the information on content, the less arbitrary should be the subsequent decisions as to what to include in the writing of any version of the test. There is a danger, however, that in the desire to be highly specific, we may go beyond our current understanding of what the components of language ability are and what their relationship is to each other. For instance, while we may believe that many sub-skills contribute to the ability to read lengthy prose passages with full understanding, it seems hardly possible in our present state of knowledge to name them all or to assess their individual contributions to the more general ability. We cannot be sure that the sum of the parts that we test will amount to the whole in which we are generally most directly interested. At the same time, however, teaching practice often assumes some such knowledge, with one sub-skill being taught at a time. It seems to us that the safest procedure is to include in the content specifications only those elements whose contribution is fairly well established.

The way in which content is described will vary with its nature. The content of a grammar test, for example, may simply list all the relevant structures and the way in which they are used in communication. The content of a test of a language skill, on the other hand, may be specified along a number of dimensions.

The description of content will also vary with the model of language and language use which we espouse. It is beyond the scope of this book to argue for any particular linguistic model. What we have done in this and subsequent chapters is to present test content in a form that has proved useful in our experience. We would not discourage readers from using other models. But whatever the model, content should be specified as fully as possible.

The following provides a framework for specifying content.



FRAMEWORK FOR SPECIFYING CONTENT

Operations (the tasks that candidates have to be able to carry out). For a reading test these might include, for example: scan text to locate specific information; guess meaning of unknown words from context.

Types of text For a writing test these might include: letters, forms, academic essays up to three pages in length.

Addressees of texts This refers to the kinds of people that the candidate is expected to be able to write or speak to (for example, expert speakers of the same age and status); or the people for whom reading and listening materials are primarily intended (for example, expert speaker university students).

Length of text(s) For a reading test, this would be the length of the passages on which items are set. For a listening test it could be the length of the spoken texts. For a writing test, the length of the pieces to be written.

Topics Topics may be specified quite loosely and selected according to suitability for the candidate and the type of test.

Readability Reading passages may be specified as being within a certain range of readability³.

Structural range Either: (a) a list of structures which may occur in texts, together with their functions

or (b) a list of structures which should be excluded

or (c) a general indication of range of structures (e.g. in terms of frequency of occurrence in the language).

Vocabulary range This may be loosely or closely specified. Examples of the latter are to be found in the specifications for the Cambridge English examinations at lower levels (such as *KET*), for each of which a word list is provided.

Dialect, accent, style This may refer to the dialects and accents that test-takers are meant to understand or those in which they are expected to write or speak. Style may be formal, informal, conversational, etc.

Speed of processing For reading this may be expressed in the number of words to be read per minute (and will vary according to type of reading to be done). For speaking it will be rate of speech, also expressed in words per minute. For listening it will be the speed at which texts are spoken.

ii. Structure, timing, medium/channel and techniques

The following should be specified:

Test structure What sections will the test have and what will be tested in each? (for example: three sections – grammar, careful reading, expeditious reading)

Number of items (in total and in the various sections)

Number of passages (and number of items associated with each)

Medium/channel (paper and pencil, tape, computer, face-to-face, telephone, etc.)

Timing (for each section and for entire test)

Techniques What techniques will be used to measure what skills or sub-skills?

³ The *Flesch Reading Ease Score* and the *Flesch–Kincaid Grade Level Score* are readily available for any passage in Microsoft Word. These measures are based on average sentence length and the average number of syllables per word. While they may not be wholly valid measures, they are at least objective.

iii. Criterial levels of performance

The required level(s) of performance for (different levels of) success should be specified. This may involve a simple statement to the effect that, to demonstrate 'mastery', 80 percent of the items must be responded to correctly.

For speaking or writing, however, one can expect a description of the criterial level to be more complex. The following is an invented example of criterial levels that might be set for an English speaking test for bank employees.

Accuracy Pronunciation must not interfere with intelligibility, even if influenced by the L1. Some errors of grammar are acceptable if they do not significantly affect meaning. The number of errors should not be so high that they become a source of irritation to the listener. Errors of vocabulary should not cause misunderstandings.

Appropriacy The use of language must be appropriate to interaction with clients and counterparts in other banks.

Range The candidate must have sufficient range of language so that s/he does not have to break everything down to a series of basic utterances. Range should be sufficient in order to follow clearly pronounced utterances on subjects appropriate to banking, and in everyday social exchanges.

Flexibility In managing interactions, the candidate must be able to initiate and close topics, repairing any breakdowns in communication that may occur.

iv. Scoring procedures

These are always important, but particularly so where scoring will be subjective. The test developers should be clear as to how they will achieve high reliability and validity in scoring. What rating scale will be used? How many people will rate each piece of work? What happens if two or more raters disagree about a piece of work?

3. Writing and moderating items

Once specifications are in place, the writing of items can begin.

i. Sampling

It is most unlikely that everything found under the heading of 'Content' in the specifications can be covered by the items in any one version of the test. Choices have to be made. For content validity and for beneficial backwash, the important thing is to choose widely from the whole area of content. One should not concentrate on those elements known to be easy to test. Succeeding versions of the test should also sample widely and unpredictably, although one will always wish to include elements that are particularly important.

ii. Writing items

Items should always be written with the specifications in mind. It is no use writing 'good' items if they are not consistent with the specifications. As one writes an item, it is essential to try to look at it through the eyes of test-takers and imagine how they might misinterpret the item (in which case it will need to be rewritten). Even if there is no possibility of misinterpretation, test-takers (especially intelligent ones) may find responses that are different from, but equally valid as, the one intended. Mention of the intended response is a reminder that the key to an item (i.e. a statement of the correct response or responses) is an integral part of the item. An item without a key is incomplete.

The writing of successful items (in the broadest sense, including, for example, the setting of writing tasks) is extremely difficult. No one can expect to be able consistently to produce perfect items. Some items will have to be rejected, others reworked. For this reason, more items should be written than the number specified for the test. It is not unusual for up to a third of multiple choice items to be rejected. The best way to identify items that have to be improved or abandoned is through the process of moderation.

iii. Moderating items

Moderation is the scrutiny of proposed items by (ideally) at least two colleagues, neither of whom is the author of the items being examined. Their task is to try to find weaknesses in the items and, where possible, remedy them. Where successful modification is not possible, they must reject the item. It is to be hoped, of course, that they will not find fault with most of the items that they moderate and that they can therefore accept them. A checklist of the kind in Table 2 (designed for moderating grammar items) is useful to moderators.

TABLE 2: MODERATION OF GRAMMAR ITEMS		
	YES	NO
1. Is the English grammatically correct?		
2. Is the English natural and acceptable?		
3. Is the English in accordance with the specifications?		
4. Does the item test what it is supposed to test, as specified?		
5. The correct response cannot be obtained without the appropriate knowledge of grammar (other than by random guessing)		
6. Is the item economical?		
7. a. Multiple choice – is there just one correct response? b. Gap filling – are there just one or two correct responses?		
8. Multiple choice: Are all the distractors likely to distract?		
9. Is the key complete and correct?		

4. Informal trialling of items on expert speakers

Items which have been through the process of moderation should be presented in the form of a test (or tests) to a number of expert speakers – twenty or more, if possible. There is no need to do this formally; the ‘test’ can be taken in the participants’ own time. The expert speakers should be similar to the people for whom the test is being developed, in terms of age, education and general background. There is no need for them to be specialists in language or testing. Indeed, it is preferable that they should not be, since ‘experts’ are unlikely to behave in the same way as naïve test-takers.

Items that prove difficult for the expert speakers almost certainly need revision or replacement. So do items where unexpected or inappropriate responses are provided. Of course, people taking a test on their own language will have lapses of attention. Where these can be recognised, the responses should not count against the item.

5. Trialling of the test on a group of non-expert speakers similar to those for whom the test is intended

Those items that have survived moderation and informal trialling on expert speakers should be put together into a test, which is then administered under test conditions to a group similar to that for which the test is intended⁴. Problems in administration and scoring are noted.

It has to be accepted that, for a number of reasons, trialling of this kind is often not feasible. In some situations a group for trialling may simply not be available. In other situations, although a suitable group exists, it may be thought that the security of the test might be put at risk. It is often the case, therefore, that faults in a test are discovered only after it has been administered to the target group. Unless it is intended that no part of the test should be used again, it is worthwhile noting problems that become apparent during administration and scoring, and afterwards carrying out statistical analysis of the kind referred to below and treated more fully in Chapter 19.

⁴ If there are too many items for one group to take in a single sitting, more than one form of the test can be constructed, with each form containing a subset of items common to both (known as anchor items). Using performance on the common anchor items as a basis for comparison, it is possible to put the other items on the same difficulty scale. If this is not done, differences in ability between the groups will mean that the difficulty levels of items taken by one group will not be directly comparable with the difficulty levels of items taken by another group. See Chapter 19 for statistical treatment of results when anchor items are used.

6. Analysis of results of the trial; making of any necessary changes

There are two kinds of analysis that should be carried out. The first – statistical – is described in Chapter 19. This will reveal qualities (such as reliability) of the test as a whole and of individual items (for example, how difficult they are, how well they discriminate between stronger and weaker candidates).

The second kind of analysis is qualitative. Responses should be examined in order to discover misinterpretations, unanticipated but possibly correct responses, and any other indicators of faulty items. Items that analysis shows to be faulty should be modified or dropped from the test. Assuming that more items have been trialled than are needed for the final test, a final selection can be made, basing decisions on the results of the analyses.

7. Calibration of rating scales

Where rating scales are going to be used for oral testing or the testing of writing, these should be calibrated. Essentially, this means collecting samples of performance (for example, pieces of writing) which cover the full range of the scales. A team of 'experts' then looks at these samples and assigns each of them to a point on the relevant scale. The assigned samples provide reference points for all future uses of the scale, as well as being essential training materials. If necessary, the scales may be modified to take account of features in the samples which they currently fail to capture.

8. Validation

The final version of the test can be validated. For a high-stakes or published test, this should be regarded as essential. For relatively low-stakes tests that are to be used within an institution, this may not be thought necessary, although where the test is likely to be used many times over a period of time, informal, small-scale validation is still desirable.

9. Writing handbooks for test-takers, test users and staff

Handbooks (each with rather different content, depending on audience) may be expected to contain the following:

- the rationale for the test;
- an account of how the test was developed and validated;

- a description of the test, giving details of sections, timings, etc. (which may include a version of the specifications);
- sample items (or a complete sample test);
- advice on preparing for taking the test;
- an explanation of how test scores are to be interpreted;
- training materials (for interviewers, raters, etc.);
- details of test administration.

The handbooks should be made available in print form or/and online.

10. Training staff

Using the handbook and other materials, all staff who will be involved in the test process should be trained. This may include interviewers, raters, scorers, computer operators and invigilators (proctors).

11. Test maintenance

If a test is to be used repeatedly over time, statistical and qualitative analysis should be carried out regularly in order to identify any problems that may have crept in. At some point, alternative versions are likely to become necessary, as word spreads of the original test's content. In this case, the development process will have to be repeated, beginning with the writing of items (assuming there is no perceived need to change the specifications).

Two examples of test development follow.

EXAMPLE OF TEST DEVELOPMENT 1: AN ACHIEVEMENT TEST

Statement of the problem

There is a need for an achievement test to be administered at the end of a pre-session course of training in the reading of academic texts in the social sciences and business studies (the students are graduates who are about to follow postgraduate courses in English-medium universities). The teaching institution concerned (as well as the sponsors of the students) wants to know just what progress is being made during the three-month course. The test must therefore be sufficiently sensitive to measure gain over that relatively short

period. While there is no call for diagnostic information on individuals, it would be useful to know, for groups, where the greatest difficulties remain at the end of the course, so that future courses may give more attention to these areas. Backwash is considered important; the test should encourage the practice of the reading skills that the students will need in their university studies. This is, in fact, intended to be only one of a battery of tests, and a maximum of two hours can be allowed for it. It will not be possible at the outset to write separate tests for different subject areas.

Specifications

Content

Operations These are based on the stated objectives of the course, and include expeditious and slower, careful reading.

Expeditious reading: Skim for main ideas; search read for information; scan to find specific items in lists, indexes, etc.

Slower, careful reading: Construe the meaning of complex, closely argued passages.

Underlying skills that are given particular attention in the course:

- Guessing the meaning of unfamiliar words from context;
- Identifying referents of pronouns etc. often some distance removed in the text.

Types of text The texts should be authentic, academic (taken from textbooks and journal articles).

Addressees Academics at postgraduate level and beyond.

Lengths of texts Expeditious: c. 3,000 words Careful: c. 800 words.

Topics The subject areas will have to be as 'neutral' as possible, since the students are from a variety of social science and business disciplines (economics, sociology, management etc.).

Readability Not specified.

Structural range Unlimited.

Vocabulary range General academic, not specialist technical.

Dialect and style Standard American or British English dialect. Formal, academic style.

Speed of processing Expeditious: 300 words per minute (not reading all words).
Careful: 100 words per minute.

Structure, timing, medium and techniques

Test structure Two sections: expeditious reading; careful reading.

Number of items 30 expeditious; 20 careful. Total: 50 items.

Number of passages 3 expeditious; 2 careful.

Timing Expeditious: 15 minutes per passage (each passage collected after 15 minutes).

Careful: 30 minutes (passage only handed out after 45 minutes, when expeditious reading has been completed).

TOTAL: 75 minutes.

Medium Paper-and-pencil. Each passage in a separate booklet.

Techniques Short answer and gap filling for both sections.

Examples:

- a) For inferring meaning from context:

For each of the following, find a single word in the text with an equivalent meaning. Note: the word in the text may have an ending such as *-ing*, *-s*, etc.

highest point (lines 20–35)

- b) For identifying referents:

What does each of the following refer to in the text? Be very precise.

the former (line 43)

Criterial levels of performance

Satisfactory performance is represented by 80 percent accuracy in each of the two sections.

The number of students reaching this level will be the number who have succeeded in terms of the course's objectives.

Scoring procedures

There will be independent double scoring. Scorers will be trained to ignore irrelevant (for example, grammatical) inaccuracy in responses.

Sampling

Texts will be chosen from as wide a range of topics and types of writing as is compatible with the specifications. Draft items will only be written after the suitability of the texts has been agreed.

Item writing and moderation

Items will be based on a consideration of what a competent non-specialist reader should be able to obtain from the texts. Considerable time will be set aside for moderation and rewriting of items.

Informal trialling

This will be carried out on 20 expert speaker postgraduate students in the university.

Trialling and analysis

Trialling of texts and items sufficient for at least two versions will be carried out with students currently taking the course, with full qualitative and statistical

analysis. An overall reliability coefficient of 0.90 and a percent agreement (see Chapter 5) of 0.85 are required.

Validation

There will be immediate content validation carried out by staff experienced in teaching and testing.

Concurrent validation will be against tutors' ratings of the students.

Predictive validation will be against subject supervisors' ratings one month after the students begin their postgraduate studies.

Handbooks

One handbook will be written for the students, their sponsors, and their future supervisors.

Another handbook will be written for internal use.

EXAMPLE OF TEST DEVELOPMENT 2: A PLACEMENT TEST

Statement of the problem

A commercial English language teaching organisation (which has a number of schools) needs a placement test. Its purpose will be to assign new students to classes at five levels: false beginners; lower intermediate; middle intermediate; upper intermediate; advanced. Course objectives at all levels are expressed in rather general 'communicative' terms, with no one skill being given greater attention than any other. As well as information on overall ability in the language, some indication of oral ability would be useful. Sufficient accuracy is required for there to be little need for changes of class once teaching is under way. Backwash is not a serious consideration. More than two thousand new students enrol within a matter of days. The test must be brief (not more than 45 minutes in length), quick and easy to administer, score and interpret. Scoring by clerical staff should be possible. The organisation has previously conducted interviews but the number of students now entering the school is making this impossible.

Specifications

Content

Operations Ability to predict missing words (based on the notion of 'reduced redundancy'⁵).

Length of text One turn (of a maximum of about 20 words) per person.

Types of text Constructed 'spoken' exchanges involving two people. It is hoped that the spoken nature of the texts will, however indirectly, draw on students' oral abilities.

⁵. See Chapter 14 for a discussion of reduced redundancy.

Topics 'Everyday'. Those found in the textbooks used by the organisation.

Structural range All those found in the textbooks (listed in the specifications but omitted here to save space).

Vocabulary range As found in the textbooks, plus any other common lexis.

Dialect and style Standard English English. Mostly informal style, some formal.

Structure, timing, medium and techniques

Test structure No separate sections.

Number of items 100 (though this will be reduced if the test is shown to do its job well with fewer items).

Timing 30 minutes (Note: this seems very little time, but the more advanced students will find the early passages extremely easy, and will take very little time. It does not matter whether lower-level students reach the later passages.)

Medium Pencil-and-paper.

Technique All items will be gap filling. One word per gap. Contractions count as one word. Gaps will relate to vocabulary as well as structure (not always possible to distinguish what is being tested).

Examples: A: Whose book _____ that?

B: It's mine.

A: How did you learn French?

B: I just picked it _____ as I went along.

Criteria levels of performance

These will only be decided when comparison is made between performance on the test and (a) the current assignment of students by the interview and (b) the teachers' view of each student's suitability to the class they have been assigned to by the interview.

Scoring procedures

Responses will be on a separate response sheet. A template with a key will be constructed so that scoring can be done rapidly by clerical staff.

Informal trialling

This will be carried out on 20 first-year expert speaker undergraduate students.

Trialling and analysis

Many more items will be constructed than will finally be used. All of them (in as many as three different test forms, with linking anchor items) will be trialled on current students at all levels in the organisation. Problems in administration and scoring will be noted.

After statistical and qualitative analysis, one test form made up of the 'best' items will be constructed and trialled on a different set of current students. The total score for each of the students will then be compared with his or her level in the institution, and decisions as to criterial levels of performance made.

Validation

The final version of the test will be checked against the list of structures in the specifications. If one is honest, however, one must say that at this stage content validity will be only a matter of academic interest. What will matter is whether the test does the job it is intended for. Thus the most important form of validation will be criterion-related, the criterion being placement of students in appropriate classes, as judged by their teachers (and possibly by the students themselves). The smaller the proportion of misplacements, the more valid the test.

Handbook

A handbook will be written for distribution by the organisation to its various schools.



READER ACTIVITIES

On the basis of experience or intuition, try to write a specification for a test designed to measure the level of language proficiency of students applying to study an academic subject in the medium of a foreign language at an overseas university. Compare your specification with those of tests that have actually been constructed for that purpose.



FURTHER READING

Test development process

O'Sullivan (2012b) presents an outline of the test development process. Davidson and Fulcher (2012) offer advice on the development of test specifications. Specifications for a test designed to assess the level of English of students wishing to study at tertiary level in the UK, the *Test of English for Educational Purposes (TEEP)*, are to be found in Weir (1988, 1990).

For other models of test development see Alderson et al. (1995) and Bachman and Palmer (1996). The model used by Bachman and Palmer is highly detailed and complex but their book gives information on ten test development projects.

Alderson and Buck (1993) report on the test development procedures of certain British testing bodies.

Common European Framework

Language Testing 22, 3 (2005) includes a number of articles about the use of the *Common European Framework* (see Online resources, below) in language testing.

Contribution of teachers

Cumming et al. (2004) report on the use of experienced teachers in investigating the content validity of a new test.

Handbooks

For advice on what to include in handbooks, see *AERA* (1999), which is reviewed by Davidson (2000).

Online resources

Cambridge Assessment English is a valuable source of information and examples which will help in the development of a new test of English.

ALTE (Association of Language Testers in Europe) provides advice on test development, including a variety of checklists helpful for ensuring content validity, etc.

The *COBUILD* corpus and the *British National Corpus* between them provide millions of utterances in English, which can be used as the basis for items.

The *Common European Framework of Reference for Languages (CEFR)* describes language activities and competences at six levels. Many commercial tests are linked to these levels. This is also increasingly the case for teacher-made tests.

English Profile relates grammatical structures and vocabulary items to the different *CEFR* levels, and is very useful for the development of teacher-made tests.

The *Oxford 3000™* gives what language experts and experienced teachers believe to be the 3,000 most important words for learners of English.

The ACTFL (American Council on the Teaching of Foreign Languages) website provides access to downloadable proficiency guidelines and can-do statements for numerous languages which are potentially useful in establishing test content and creating rating scales.

8 Common test techniques

What are test techniques¹?

Quite simply, test techniques are means of eliciting behaviour from candidates that will tell us about their language abilities. What we need are techniques that:

- will elicit behaviour which is a reliable and valid indicator of the ability in which we are interested;
- will elicit behaviour which can be reliably scored;
- are as economical of time and effort as possible;
- will have a beneficial backwash effect, where this is relevant.

From Chapter 9 to Chapter 13, techniques are discussed in relation to particular abilities. Techniques that may be thought to test 'overall ability' are treated in Chapter 14. The present chapter introduces common techniques that can be used to test a variety of abilities, including reading, listening, grammar and vocabulary. This is to avoid having to introduce these techniques repeatedly in the chapters in which they appear later. We begin with an examination of the multiple choice technique and then go on to look at techniques that require the test-taker to construct a response (rather than just select one from a number provided by the test-maker).

Multiple choice items

Multiple choice items take many forms, but their basic structure is as follows.

There is a *stem*:

Ashley has been here _____ half an hour.

and a number of *options* – one of which is correct, the others being *distractors*:

A. during B. for C. while D. since

It is the candidate's task to identify the correct or most appropriate option (in this case B). Perhaps the most obvious advantage of multiple choice,

¹ Test techniques are frequently referred to as 'formats'. We prefer the word 'technique', leaving the word 'format' for more general aspects of test structure, such as the interview.

referred to earlier in the book, is that scoring can be perfectly reliable. Scoring should also be rapid and economical. A further considerable advantage is that, since in order to respond the candidate has only to make a mark on the paper or, on a computer, choose from a drop-down menu, it is possible to include more items than would otherwise be possible in a given period of time. As we know from Chapter 5, this is likely to make for greater test reliability. Finally, it allows the testing of receptive skills without requiring the test-taker to produce written or spoken language.

The advantages of the multiple choice technique were so highly regarded at one time that it almost seemed that it was the only way to test. While many laymen have always been sceptical of what could be achieved through multiple choice testing, it is only fairly recently that the technique's limitations have been more generally recognised by professional testers. The difficulties with multiple choice are as follows.

The technique tests only recognition knowledge

If there is a lack of fit between at least some candidates' productive and receptive skills, then performance on a multiple choice test may give a quite inaccurate picture of those candidates' ability. A multiple choice grammar test score, for example, may be a poor indicator of someone's ability to use grammatical structures. The person who can identify the correct response in the item above may not be able to produce the correct form when speaking or writing. This is in part a question of construct validity; whether or not grammatical knowledge of the kind that can be demonstrated in a multiple choice test underlies the productive use of grammar. Even if it does, there is still a gap to be bridged between knowledge and use; if use is what we are interested in, that gap will mean that test scores are at best giving incomplete information.

Guessing may have a considerable but unknowable effect on test scores

The chance of guessing the correct answer in a three-option multiple choice item is one in three, or roughly 33 percent. On average we would expect someone to score 33 on a 100-item test purely by guess-work. We would expect some people to score fewer than that by guessing, others to score more. The trouble is that we can never know what part of any particular individual's score has come about through guessing. Attempts are sometimes made to estimate the contribution of guessing by assuming that all incorrect responses are the result of guessing, and by further assuming that the individual has had average luck in guessing. Scores are then reduced by the number of points the individual is estimated to have obtained by guessing. However, neither assumption is necessarily correct, and we cannot know that the revised score is the same as (or very close to) the one an individual would have obtained without guessing. While other testing methods may also involve guessing, we would normally expect the

effect to be much less, since candidates will usually not have a restricted number of responses presented to them (with the information that one of them is correct).

If multiple choice is to be used, every effort should be made to have at least four options (in order to reduce the effect of guessing). It is important that all of the distractors should be chosen by a significant number of test-takers who do not have the knowledge or ability being tested. If there are four options but only a very small proportion of candidates choose one of the distractors, the item is effectively only a three-option item.

Successful guessing can be reduced by using items with five options, of which two correct answers are to be chosen by test-takers. For example:

If I had chosen a different career, _____ more money.

- a. I've made
- b. I'd have made
- c. I'll be making
- d. I'd be making
- e. I'm making

The item above is only marked as correct if the test-taker chooses both correct options (in this example, options b and d). Since, logically, guessing will be less effective than if only one correct option needs to be identified, this type of item would appear to have more validity than a traditional item with only one correct option. A drawback to this technique, though, is that items with two correct options can be more difficult to write and indeed, depending on the language point being tested, will sometimes be impossible to create.

The technique severely restricts what can be tested

The basic problem here is that multiple choice items require distractors, and distractors are not always available. In a grammar test, it may not be possible to find three or four plausible alternatives to the correct structure. The result is often that the command of what may be an important structure is simply not tested. An example would be the distinction in English between the past simple and the present perfect. For learners at a certain level of ability, in a given linguistic context, there are no other alternatives that are likely to distract. The argument that this must be a difficulty for any item that attempts to test for this distinction is difficult to sustain, since other items that do not overtly present a choice may elicit the candidate's usual behaviour, without the candidate resorting to guessing. In other words, 'constructed response items', where students are required to supply their own answer, allow for a greater range of structures to be tested.

It is very difficult to write successful items

A further problem with multiple choice is that, even where items are possible, good ones are extremely difficult to write. Professional test writers reckon to have to write many more multiple choice items than they actually need for a test, and it is only after trialling and statistical analysis of performance on the items that they can recognise the ones that are usable. It is our experience that multiple choice tests that are produced for use within institutions are often shot through with faults. Common amongst these are: more than one correct answer; no correct answer; there are clues in the options as to which is correct (for example, the correct option may be different in length from the others); ineffective distractors. The amount of work and expertise needed to prepare good multiple choice tests is so great that, even if one ignored other problems associated with the technique, one would not wish to recommend it for regular achievement testing (where the same test is not used repeatedly) within institutions. Savings in time for administration and scoring will be outweighed by the time spent on successful test preparation. It is true that the development and use of item banks, from which a selection can be made for particular versions of a test, makes the effort more worthwhile, but great demands are still made on time and expertise.

Backwash may be harmful

It should hardly be necessary to point out that where a test that is important to students is multiple choice in nature, there is a danger that practice for the test will have a harmful effect on learning and teaching. Practice at multiple choice items (especially when – as can happen – as much attention is paid to improving one's educated guessing as to the content of the items) will not usually be the best way for students to improve their command of a language.

Cheating may be facilitated

The fact that the responses on a multiple choice test (a, b, c, d) are so simple makes them easy to communicate to other candidates non-verbally. Some defence against this is to have at least two versions of the test, the only difference between them being the order in which the options are presented.

All in all, the multiple choice technique is best suited to relatively infrequent testing of large numbers of candidates. This is not to say that there should be no multiple choice items in tests produced regularly within institutions. In setting a reading comprehension test, for example, there may be certain tasks that lend themselves very readily to the multiple choice format, with obvious distractors presenting themselves in the text. There are real-life tasks (say, a shop assistant identifying which one of four dresses a customer is describing) which are essentially multiple choice. The simulation in a test of such a situation would seem to be perfectly appropriate. What the reader is being urged to avoid is the excessive, indiscriminate and potentially harmful use of the technique.

Having identified problems with multiple choice items, we have to recognise that teachers are often required to write them. In Chapters 11, 12, 13 and 15, advice is given on writing multiple choice items for particular purposes. In the meantime, with this in mind, we include here a set of guidelines to help avoid the most common pitfalls. Teachers can use this as a checklist, while always bearing in mind the various issues with this technique as described in this chapter.



GUIDELINES FOR WRITING EFFECTIVE MULTIPLE CHOICE ITEMS

1. Include at least four options for each item.
2. Keep all options a similar length to each other.
3. Vary where the correct option comes in each item (e.g. option d should not be the correct option more often than a, b or c).
4. Make sure all distractors are plausible. Consider using students' incorrect answers given in previous constructed response tests.
5. Make sure none of the distractors are possible as correct answers.
6. Don't try to trick test-takers.
7. Include the majority of the words in the stem and keep the options short.
8. Always ask your peers to check the items as if they were taking the test. Then edit where necessary based on any issues identified by your peers.

Yes/No and True/False items

Items in which the test-taker has merely to choose between *Yes* and *No*, or between *True* and *False*, are effectively multiple choice items with only two options. The attraction of this technique is the speed at which they can be written and answered. However, the obvious weakness of such items is that the test-taker has a 50 percent chance of choosing the correct response by chance alone². In our view, there is no place for items of this kind in a formal test, although they may well have a use as part of informal, formative assessment where the accuracy of the results is not critical. *True/False* items are sometimes modified by requiring test-takers to give a reason for their choice. However, this extra requirement is problematic, first because it is adding what is a potentially difficult writing task when writing is not meant to be tested (validity problem), and secondly because the responses are often difficult to score (reliability and validity problem). Items of this kind may be improved slightly by requiring candidates to justify their choice of *Yes* or *No* by identifying a phrase or sentence in the text which supports their choice (by underlining or copying). This clearly removes the potentially difficult writing task, but in practice it is often difficult to specify all acceptable responses. For example, there may be more than one sentence offering support.

². This can be improved slightly with items that have three options (*true/false/doesn't say*).

Short-answer items

Items in which the test-taker has to provide a short answer are common, particularly in listening and reading tests.

Examples:

- i. What does *it* in the last sentence refer to?
- ii. How old was Harry Potter when he started doing magic?
- iii. Why was Harry unhappy?

Advantages of short-answer items over multiple choice are that:

- guessing will (or should) contribute less to test scores;
- the technique is not restricted by the need for distractors (though there have to be potential alternative responses);
- cheating is likely to be more difficult;
- though great care must still be taken, items should be easier to write.

Disadvantages are:

- responses may take longer and so reduce the possible number of items, which in turn has the potential to reduce the test's reliability;
- the test-taker has to produce language in order to respond;
- scoring may be invalid or unreliable, if judgement is required;
- scoring may take longer.

The first two of these disadvantages may not be significant if the required response is really short (and at least the test-takers do not have to ponder four options, three of which have been designed to distract them). The next two can be overcome by making the required response unique (i.e. there is only one possible answer) and to be found in the text (or to require very simple language). Looking at the examples above, without needing to see the text, we can see that the correct response to Item i. should be unique and found in the text. The same could be true of Item ii. Item iii., however, may cause problems (which can be solved by using gap filling, below).

We believe that short-answer questions have a role to play in serious language testing. Only when testing has to be carried out on a very large scale would we think of dismissing short-answer questions as a possible technique because of the time taken to score. With the increased use of computers in testing (in *TOEFL*[®], for example), where written responses can be scored reliably and quickly, there is no reason for short-answer items not to have a place in the very largest testing programmes.

Gap filling items

Items in which test-takers have to fill a gap with a word are also common. An example for a reading test might be:

Harry was unhappy because his parents _____ when he was young and he was _____ at school.

From this example, assuming that the missing words (let us say they are *died* and *bullied*) can be found in the text, it can be seen that the problem of the third short-answer item has been overcome. Gap filling items for reading or listening work best if the missing words are to be found in the text or are straightforward, high frequency words which should not present spelling problems.

Gap filling items can also work well in tests of grammar and vocabulary. Examples:

He asked me for money, _____ though he knows I earn a lot less than him.

Our son just failed another exam. He really needs to pull his _____ up.

But it does not work well where the grammatical element to be tested is discontinuous, and so needs more than one gap. An example would be where one wants to see if the test-taker can provide the past continuous appropriately. None of the following is satisfactory:

- i. While they _____ watching television, there was a sudden bang outside.
- ii. While they were _____ television, there was a sudden bang outside.
- iii. While they _____ television, there was a sudden bang outside.

In the first two cases, alternative structures which the test-taker might have naturally used (such as the simple past) are excluded. The same is true in the third case too, unless the test-taker inserted an adverb and wrote, for example, *quietly watched*, which is an unlikely response. In all three cases, there is too strong a clue as to the structure which is needed.

Gap filling does not always work well for grammar or vocabulary items where minor or subtle differences of meaning are concerned, as the following items demonstrate.

- i. A: What will he do?
B: I think he _____ resign.

A variety of modal verbs (*will, may, might, could*, etc.) can fill the gap satisfactorily.

Providing context can help:

- ii. A: I wonder who that is.
B: It _____ be the doctor.

This item has the same problem as the previous one. But adding:

A: How can you be so certain?

means that the gap must be filled with a modal expressing certainty (*must*). But even with the added context, *will* may be another possibility.

When the gap filling technique is used, it is essential that test-takers are told very clearly and firmly that only one word can be put in each gap. They should also be told whether contractions (*I'm, isn't, it's*, etc.) count as one word. This is particularly important if the test is to be computer marked, as there will be no possibility of marker discretion. (In our experience, counting contractions as one word is advisable, as it allows greater flexibility in item construction.)

Gap filling is a valuable technique. It has the advantages of the short-answer technique, but the greater control it exercises over the test-takers means that it does not call for significant productive skills. There is no reason why the scoring of gap filling should not be highly reliable, provided that it is carried out with a carefully constructed key on which the scorers can rely completely (and not have to use their individual judgement).

One recent development is the use of corpora and computer algorithms to assist with gap filling item writing. Some programs create items based on a keyword which a user submits, while others will take a text and automatically replace certain words with gaps. The choice of which words are to be gapped is of course crucial. One program run by the University of Nottingham chooses words based on different levels of the *Academic Word List*, thereby allowing users to vary the difficulty of the task. While these programs undoubtedly have the potential to be useful tools, their output needs to be scrutinised and modified where necessary before use. However, as algorithms continue to be developed and finely tuned, it will be interesting to see to what extent they can replace human item writers.

This chapter has only provided an introduction to certain common testing techniques. The techniques are treated in greater detail in later chapters, along with others that are relevant to the testing of particular abilities.



READER ACTIVITIES

1. Examine each of the following three items. If an item is problematic, what is the problem? Can you remove the problem without changing the technique?

- i. When she asked for an extension, they agreed _____ let her have another month to finish the report.
a. at b. to c. over d. of
Key: b
 - ii. A: Why are you doing the work yourself?
B: When I asked Bill, he said he _____ do it.
Key: couldn't
 - iii. A: It's too easy for young people to make money these days.
B: I _____ agree more.
Key: couldn't
2. Rewrite each of the above items using another technique. What do you learn from doing this?
 3. Look at ten items in any test to which you have access. If any of them are problematic, can you improve them using the same technique as in the original item? See how many of the ten items can be satisfactorily rewritten using a different technique.
 4. Visit the University of Nottingham AWL gapmaker site (search terms 'AWL gapmaker Nottingham') and submit a text of between 200 and 300 words. Select different sublists to be used and notice how this affects the gap filling task. Do any of the sublists generate a task that is suitable for your students?
 5. Try writing a multiple choice item with two correct answers and three distractors. Show your item to a colleague and ask them to evaluate it. How easy do you find it to write an item like this? What was the biggest challenge in writing this item?
 6. Find an element for which you cannot successfully construct an item with two correct responses and three distractors. Challenge a colleague to write one on the same element.



FURTHER READING

Heaton (1975) discusses various types of item and gives many examples for analysis by the reader. Amini and Ibrahim-González (2012) suggest the backwash effects of the multiple choice technique are not as beneficial as the cloze technique. Their study focuses specifically on vocabulary acquisition. Currie and Chiraramanee's research (2010) casts further doubt on the validity of the multiple choice technique, particularly in comparison to constructed response items. Smith et al. (2010) gives a detailed description and evaluation of a corpus-driven gap filling system, *TEDDCLOG*.

9

Testing writing

We will make the assumption in this chapter that the best way to test people's writing ability is to get them to write¹.

Given the decision to test writing ability directly, we are in a position to state the testing problem, in a general form, for writing. This has three parts:

1. We have to set writing tasks that are properly representative of the population of tasks that we should expect the students to be able to perform.
2. The tasks should elicit valid samples of writing (i.e. which truly represent the students' ability).
3. It is essential that the samples of writing can and will be scored validly and reliably.

We shall deal with each of these in turn, offering advice and examples.

Representative tasks

i. Specify all possible content

In order to judge whether the tasks we set are representative of the tasks that we expect students to be able to perform, we have to be clear at the outset just what these tasks are that they should be able to perform. These should be identified in the test specifications. The following elements in the framework for the specification of content presented in Chapter 7 are relevant here: operations, types of text, addressees, length of texts, topics, dialect and style.

Let us look at the writing section of the current handbook of the *Cambridge English B2 First*. The description of the Writing paper may not include the complete set of specifications for the two parts of the test but it shows what specifications for a writing test may look like.

¹ We will also assume that the writing of elementary students is not to be tested. Whatever writing skills are required of them can be assessed informally. There seems little point in constructing, for example, a formal test of the ability to form characters or transcribe simple sentences.

Operations

agreeing or disagreeing with a statement
 giving information and explanations
 giving opinions on a question
 exemplifying
 giving reasons
 comparing and contrasting ideas and opinions
 drawing a conclusion
 describing
 explaining
 reporting
 suggesting
 recommending
 persuading

Types of text

an essay, an article, an informal email or letter, a report, a review

Addressees of texts

articles for an English language magazine or newsletter
 emails/letters for (for example) friends, colleagues, potential employers, college principal, magazine editor
 essay for the teacher
 report for a teacher or a peer group
 review for magazines, websites or newspaper

Topics

'a range of topics, such as health and fitness, sport, music and so on'

Dialect and length of texts

140–190 words. Dialects are unspecified.

It is probably fair to say that the *B2 First* writing specifications (as they appear in the handbook) account for a significant proportion of the writing tasks that students in general language courses that have communicative aims are expected to be able to perform. They ought, therefore, to be useful to readers of this book who are responsible for testing writing on such courses. Under each heading, institutional testers can identify the

elements that apply to their own situation. There will be some points where perhaps more detail is called for; others where additional elements are needed. There is certainly no reason to feel limited to this particular framework or its content, but all in all these specifications should provide a good starting point for many testing purposes. For the same reason, further examples of specifications are given in the following chapters.

A second example, this time much more restricted, concerns the writing component of a test of English for academic purposes with which one of us was associated. The purpose of the test was to discover whether a student's written English was adequate for study through the medium of English at a particular overseas university. An analysis of needs had revealed that the most important uses of written English were for the purpose of taking notes in lectures and the writing of examination answers up to two paragraphs in length. The first of these tasks was integrated into the listening component of the test. This left the examination answers. An analysis of examination questions in the university revealed that students were required to describe, explain, compare and contrast, and argue for and against a position. Because in that university the first-year undergraduate course is very general (all students study arts, science and social science subjects), almost all reasonably academic topics were appropriate. The addressees were university lecturers – both expert speakers and non-expert speakers of English. Using the suggested framework, we can describe the relevant tasks quite succinctly:

Operations

Describe, explain, compare and contrast, argue for and against a position.

Types of text

Examination answers up to two paragraphs in length.

Addressees of texts

Expert speaker and non-expert speaker university lecturers.

Topics

Any capable of academic treatment. Not specialist. Relevant to the test-takers.

Dialect and style

Any standard variety of English (e.g. American, British) or a mixture of these. Formal style.

Length of texts

About one page.

ii. Include a representative sample of the specified content

From the standpoint of content validity, the ideal test would be one which required candidates to perform all the relevant potential writing tasks. The total score obtained on that test (the sum of the scores on each of the different tasks) would be our best estimate of a candidate's ability. If it were ever possible to do this, we would not expect all of a candidate's scores to be equal, even if they were perfectly scored on the same scale. People will simply be better at some tasks than others. So, if we aren't able to include every task (and of course this is normally the case) and happen to choose just the task or tasks that a candidate is particularly good (or bad) at, then the outcome is likely to be very different. This is why we try to select a representative set of tasks. And the more tasks (within reason) that we set, the more representative of a candidate's ability (the more valid) will be the totality of the samples (of the candidate's ability) we obtain. It is also to be remembered that if a test includes a wide-ranging and representative sample of specifications, the test is more likely to have a beneficial backwash effect.

Let us look at the sample below, which appears in the *Cambridge B2 First Handbook for Teachers*.

Part 1

You **must** answer this question. Write your answer in **140 – 190** words in an appropriate style on the separate answer sheet.

- 1 In your English class you have been talking about the environment. Now, your English teacher has asked you to write an essay.

Write an essay using **all** the notes and giving reasons for your point of view.

<p>Every country in the world has problems with pollution and damage to the environment. Do you think these problems can be solved?</p>
<p>Notes</p> <p>Write about:</p>
<ol style="list-style-type: none"> 1. transport 2. rivers and seas 3. (your own idea)

Part 2

Write an answer to **one** of the questions **2 – 4** in this part. Write your answer in **140 – 190** words in an appropriate style on the separate answer sheet. Put the question number in the box at the top of the answer sheet.

- 2 You see this announcement in your college English-language magazine.

Book reviews wanted

Have you read a book in which the main character behaved in a surprising way?

Write us a review of the book, explaining what the main character did and why it was surprising. Tell us whether or not you would recommend this book to other people.

The best reviews will be published in the magazine.

Write your **review**.

- 3 You see this announcement on an English-language website.

Articles wanted

The most useful thing I have ever learned.

What is the most useful thing you have learned?
Who did you learn it from? Why is it useful?

Write us an article answering these questions.

We will publish the best articles on our website.

Write your **article**.

- 4 You have received this email from your English-speaking friend David.

From: David

Subject: touring holiday

Some college friends of mine are visiting your area soon for a week's touring holiday. They would like to travel around and learn about your local area and its history.

Can you tell me about some of the places they could visit? What's the best way to travel around – car, bike or coach?

Thanks,

David

Write your **email**.

Readers may wish to refer back to the B2 test specifications on page 88. It soon becomes clear that, despite there being a total of four tasks in the sample task above, since a candidate will only complete two tasks, he or she will only be tested on a small fraction of the operations and task types specified in the handbook. Therefore, the test's content validity is inevitably brought into question. This illustrates how really good coverage of the range of potential tasks is often not possible in a single version of a test.

This is a problem to which there is no easy answer. Only research will tell us whether candidates' performance on one small set of selected tasks will result in scores very similar to those that their performance on another small, non-overlapping set would have been awarded.

In the case of the English-medium university, it is not nearly as difficult to select representative writing tasks. Content validity is less of a problem

than with the much wider-ranging *Cambridge English B2 First* test. Since it is only under the heading of 'operations' that there is any significant variability, a test that required the student to write four answers could cover the whole range of tasks, assuming that differences of topic really did not matter. In fact, the writing component of each version of the test contained two writing tasks, and so 50 percent of all tasks were to be found in each version of the test. Topics were chosen with which it was expected all students would be familiar, and information or arguments were provided (see example, page 96).

Of course, the desirability of wide sampling has to be balanced against practicality; otherwise we would always try to include all (or at least a large proportion) of the potential tasks. It must be remembered, however, that if we need to know something accurate and meaningful about a person's writing ability, we have to be prepared to pay for that information. What we decide to do will depend in large part on how accurate the information has to be. This in turn depends on how high the stakes are. If the test is used simply to place students in classes from which they can easily be moved to another more appropriate one, accuracy is not so important; we may be satisfied with a single sample of writing. But if the result is going to be very important to candidates – if it could, for example, determine whether they are allowed to study overseas – then certainly more than one sample is necessary if serious injustices are not to be perpetrated.

Elicit a valid sample of writing ability

Set as many separate tasks as is feasible

This requirement is closely related to the need to include a representative sample of the specified content. As we saw in Chapter 5, people's performance even on the same task is unlikely to be perfectly consistent. Therefore, we have to offer candidates as many 'fresh starts' as possible, and each task can represent a fresh start. By doing this, we will achieve greater reliability and so greater validity. Again, there has to be a balance between what is desirable and what is practical.

Test only writing ability, and nothing else

This advice assumes that we do not want to test anything other than the ability to write. In language testing, we are not normally interested in knowing whether students are creative, imaginative, or even intelligent, have wide general knowledge, or have good reasons for the opinions they happen to hold. Therefore, for the sake of validity, we should not set tasks which measure these abilities. Look at the following tasks which, though invented, are based on others taken from well-known tests.

1. Write the conversation you have with a friend about the holiday you plan to have together.

2. You spend a year abroad. While you are there, you are asked to talk to a group of young people about life in your country. Write down what you would say to them.
3. 'Envy is the sin which most harms the sinner.' Discuss.
4. Discuss the advantages and disadvantages of being born into a wealthy family.

The first task seems to make demands on creativity, imagination, and indeed on script-writing ability. Success at the second would seem to depend to at least some extent on the ability to give talks. It is in fact hard to imagine either of the tasks being derived from a careful specification of writing tasks. The third and fourth tasks clearly favour candidates who have, or can instantly create, an ordered set of arguments on any topic which they meet. A clear indication that not only language ability is being tested is the fact that many educated expert speakers (including us) would not be confident of completely satisfying the examiners. Francis Bacon might have done well, if his answers were not thought too brief.

Another ability that at times interferes with the accurate measurement of writing ability is that of reading. While it is perfectly acceptable to expect the candidate to be able to read simple instructions, care has to be taken to ensure these can be fully understood by everyone whose ability is of sufficiently high standard otherwise to perform adequately on the writing task. Nor should the instructions be too long. Part (b) of the following item may be thought to suffer from both these faults.

Answer **ONE** of the following questions in about **250 words**:

Either (a) You've been asked to contribute an article to an international magazine, which is running a series called "A Good Read". Write, for the magazine, a review of a book you like.

Or (b) You have recently heard that each year the Axtel Corporation offers the opportunity for a small number of people to spend between three and six months working in one of their offices in Australia, New Zealand, the United States or Britain. The aim of the scheme is to promote international understanding, and to foster an awareness of different working methods.

Candidates for the scheme are asked to write an initial letter of application, briefly outlining their general background and, more importantly, giving the reasons why they feel they would benefit from the scheme. In addition, they should indicate in which country they would like to work. On the basis of this letter they may be invited for interview and offered a post.

Write the letter of application.

One way of reducing dependence on the candidates' ability to read is to make use of illustrations.

For example:

Write one sentence based on the following picture. Include the words **so** and **difficult**.

Your sentence will be scored on:

Appropriate use of grammar and relevance to the picture.



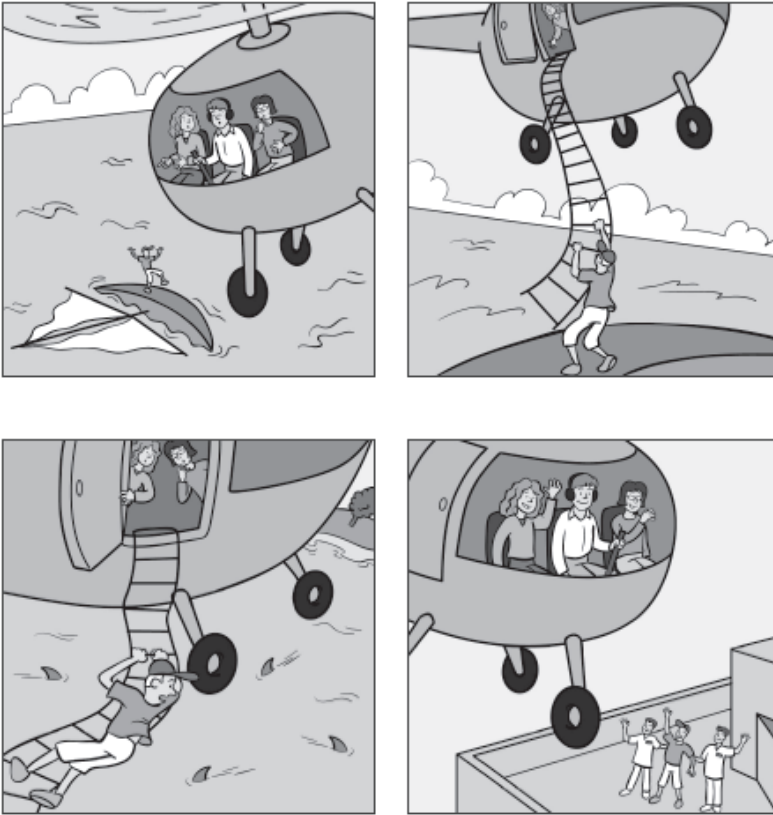
A series of pictures can be used to elicit a narrative. The following example is taken from the Breakthrough level of the *Pearson Test of English for Young Learners*.

6. Task Six: A Helicopter Ride (20 marks)

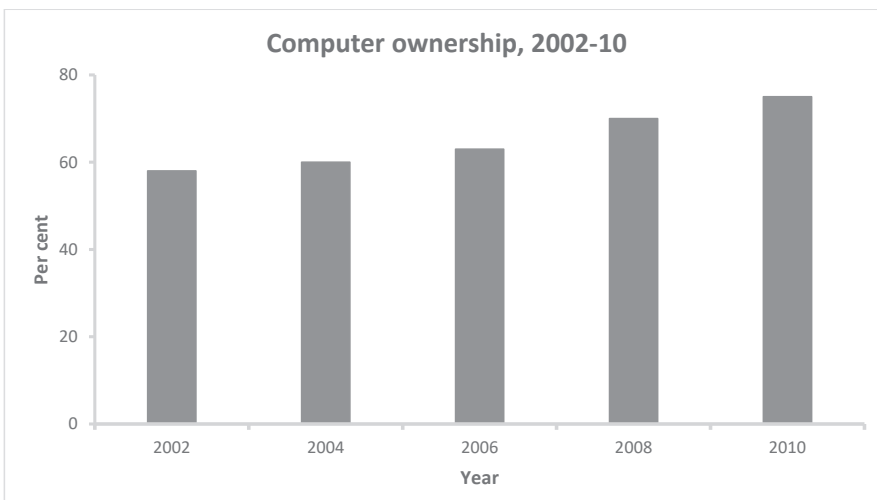
Uncle Peter takes Anna and Kirsty for a ride in his helicopter. Look at the pictures and write the story. You must use all the pictures.

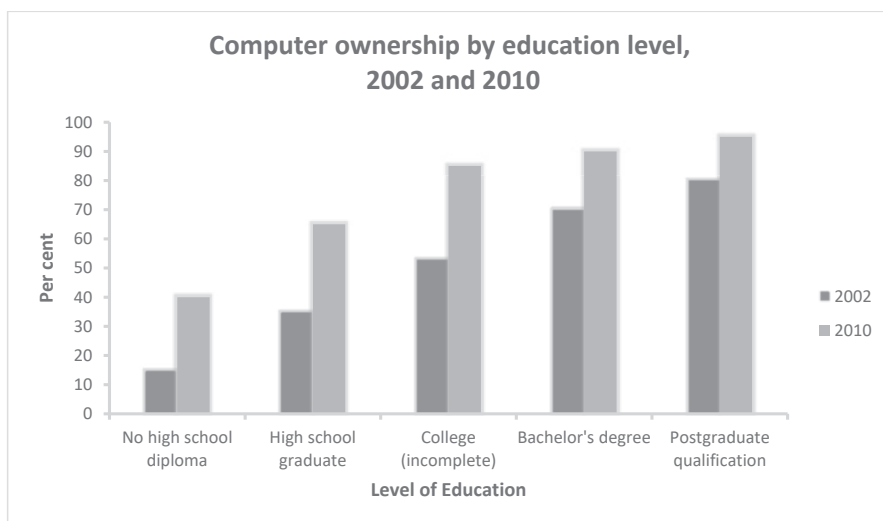
Write about 75 words.





This may take the form of a quite realistic transfer of information from graphic form to continuous prose. This example is from part one of the *IELTS Academic Writing Test*.





Restrict candidates

This echoes the general point made in Chapter 5. The question above about envy, for example, could result in very different answers from the same person on different occasions. There are so many significantly different ways of developing a response to the stimulus. Writing tasks should be well defined: candidates should know just what is required of them, and they should not be allowed to go too far astray. A useful device is to provide information in the form of notes (as in the *Cambridge B2 First* example), or a chart, as above.

The following example – slightly modified – was used in the test one of us was concerned with, mentioned earlier in the chapter.

Compare the benefits of a university education in English with that of one in Arabic. Use all of the points given below and come to a conclusion. You should write about one page.

a. Arabic

1. Easier for students
- Easier for most teachers
- Saves a year in most cases

b. English

1. Books and scientific sources mostly in English.
2. English international language – more and better job opportunities.
3. Learning second language part of education/culture.

Care has to be taken when notes are given not to provide students with too much of what they need in order to carry out the task. Full sentences are generally to be avoided, particularly where they can be incorporated into the composition with little or no change.

One last thing to say about tasks is that not only should they fit well with the specifications, but they should also be made as authentic as possible. When thinking of authenticity, it is important to take into account the nature of the candidates and their relationship with the people to or for whom the task requires them to write. A task which may be authentic for one set of candidates may be quite inauthentic for another. For example, it would be quite normal in some situations for language teachers to write to their supervisor for advice, while in other situations it would be unthinkable. While on the subject of authenticity it is worth mentioning the use of computers in writing tests. As far as possible, tests should reflect real-world writing and therefore, in many contexts, computer-based writing tests will often be more appropriate than paper-based alternatives.

Ensure valid and reliable scoring

Set tasks which can be reliably scored

A number of the suggestions made to obtain a representative performance will also facilitate reliable scoring.

Set as many tasks as possible

The more scores for each candidate, the more reliable should be the total score.

Restrict candidates

The greater the restrictions imposed on the candidates, the more directly comparable will be the performances of different candidates.

Give no choice of tasks

Making the candidates perform all tasks also makes comparisons between candidates easier.

Ensure long enough samples

The samples of writing that are elicited have to be long enough for judgements to be made reliably. This is particularly important where diagnostic information is sought. For example, in order to obtain reliable information on students' organisational ability in writing, the pieces have to be long enough for organisation to reveal itself. Given a fixed period of time for the test, there is an almost inevitable tension between the need for length and the need to have as many samples as possible.

Create appropriate scales for scoring

One expects to find the scales used in rating performance in the specifications under the heading 'criterial levels of performance'. There are two basic approaches to scoring: holistic and analytic.

Holistic scoring

Holistic scoring (sometimes referred to as 'impressionistic' scoring) involves the assignment of a single score to a piece of writing on the basis of an overall impression of it. This kind of scoring has the advantage of being very rapid. Experienced scorers can judge a one-page piece of writing in just a couple of minutes or even less (scorers of the *TOEFL® Test of Written English* apparently have just one and a half minutes for each scoring of a composition). This means that it is possible for each piece of work to be scored more than once, which is fortunate, since it is also necessary! Harris (1968) refers to research in which, when each student wrote one 20-minute composition – scored only once – the reliability coefficient was only 0.25. If well conceived and well organised, holistic scoring in which each student's work is scored by four different trained scorers can result in high scorer reliability. There is nothing magical about the number 'four'; it is simply that research has quite consistently shown acceptably high scorer reliability when writing is scored four times.

We expressed above a reservation about the need for such scoring to be well conceived. Not every scoring system will give equally valid and reliable results in every situation. The system has to be appropriate to the level of the candidates and the purpose of the test. Look at the following scoring system used in the English-medium university already referred to in this chapter.

NS	Native speaker standard
NS-	Close to native speaker standard
MA	Clearly more than adequate
MA-	Possibly more than adequate
A	ADEQUATE FOR STUDY AT THIS UNIVERSITY
D	Doubtful
NA	Clearly not adequate
FBA	Far below adequate

This scale worked perfectly well in the situation for which it was designed. The purpose of the writing component of the test was to determine whether a student's writing ability was adequate for study in English in that university. The standards set were based on an examination of undergraduate students' written work and their teachers' judgements as to the acceptability of the English therein. With students writing two compositions, each independently scored twice, using the above

scale, scorer reliability was 0.9. This is about as high as one is likely to achieve in ordinary circumstances (i.e. not in some kind of experiment or research where practicality is of no importance). It was designed for a specific purpose and obviously it would be of little use in most other circumstances. Testers have to be prepared to modify existing scales to suit their own purposes. Look now at the following, which relates to the writing component of the *TOEFL iBT*[®] (internet-based test).

Independent WRITING Rubrics

SCORE	TASK DESCRIPTION
5	<p>An essay at this level largely accomplishes all of the following:</p> <ul style="list-style-type: none"> ■ Effectively addresses the topic and task ■ Is well organized and well developed, using clearly appropriate explanations, exemplifications and/or details ■ Displays unity, progression and coherence ■ Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice and idiomaticity, though it may have minor lexical or grammatical errors
4	<p>An essay at this level largely accomplishes all of the following:</p> <ul style="list-style-type: none"> ■ Addresses the topic and task well, though some points may not be fully elaborated ■ Is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications and/or details ■ Displays unity, progression and coherence, though it may contain occasional redundancy, digression, or unclear connections ■ Displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form or use of idiomatic language that do not interfere with meaning
3	<p>An essay at this level is marked by one or more of the following:</p> <ul style="list-style-type: none"> ■ Addresses the topic and task using somewhat developed explanations, exemplifications and/or details ■ Displays unity, progression and coherence, though connection of ideas may be occasionally obscured ■ May demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning ■ May display accurate but limited range of syntactic structures and vocabulary
2	<p>An essay at this level may reveal one or more of the following weaknesses:</p> <ul style="list-style-type: none"> ■ Limited development in response to the topic and task ■ Inadequate organization or connection of ideas ■ Inappropriate or insufficient exemplifications, explanations or details to support or illustrate generalizations in response to the task ■ A noticeably inappropriate choice of words or word forms ■ An accumulation of errors in sentence structure and/or usage
1	<p>An essay at this level is seriously flawed by one or more of the following weaknesses:</p> <ul style="list-style-type: none"> ■ Serious disorganization or underdevelopment ■ Little or no detail, or irrelevant specifics, or questionable responsiveness to the task ■ Serious and frequent errors in sentence structure or usage
0	<p>An essay at this level merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.</p>

Copyright © 2014 by Educational Testing Service. All rights reserved. ETS, the ETS logo, TOEFL and TOEFL iBT are registered trademarks of Educational Testing Service (ETS) in the United States and other countries. 271.21

Though similar, this scale is different in two ways. First, because scores on the *TOEFL*[®] are used by many institutions, not just one, the headings are more general. Second, it provides some indication of the linguistic features of written work at each of the six levels. This may be useful both to the scorers and to the test score users.

If these indications become too detailed, however, a problem arises. Look at the following, which is part of the ACTFL (American Council for the Teaching of Foreign Languages) descriptors for writing, and represents an attempt to provide external criteria against which foreign language learning in schools and colleges can be assessed. The full scale has 10 points, from Novice-Low to Superior.

ADVANCED LOW

Writers at the Advanced Low sublevel are able to meet basic work and/or academic writing needs. They demonstrate the ability to narrate and describe in major time frames with some control of aspect. They are able to compose simple summaries on familiar topics. Advanced Low writers are able to combine and link sentences into texts of paragraph length and structure. Their writing, while adequate to satisfy the criteria of the Advanced level, may not be substantive. Writers at the Advanced Low sublevel demonstrate the ability to incorporate a limited number of cohesive devices, and may resort to some redundancy and awkward repetition. They rely on patterns of oral discourse and the writing style of their first language.

INTERMEDIATE HIGH

Writers at the Intermediate High sublevel are able to meet all practical writing needs of the Intermediate level. Additionally, they can write compositions and simple summaries related to work and/or school experiences. They can narrate and describe in different time frames when writing about everyday events and situations. These narrations and descriptions are often, but not always, of paragraph length, and they typically contain some evidence of breakdown in one or more features of the Advanced level. For example, these writers may be inconsistent in the use of appropriate major time markers, resulting in a loss of clarity. The vocabulary, grammar and style of Intermediate High writers essentially correspond to those of the spoken language. Intermediate High writing, even with numerous and perhaps significant errors, is generally comprehensible to natives not used to the writing of non-natives, but there are likely to be gaps in comprehension.

INTERMEDIATE MID

Writers at the Intermediate Mid sublevel are able to meet a number of practical writing needs. They can write short, simple communications, compositions, and requests for information in loosely connected texts about personal preferences, daily routines, common events, and other personal topics. Their writing is framed in present time but may contain references to other time frames. The writing style closely resembles oral discourse. Writers at the Intermediate Mid sublevel show evidence of control of basic sentence structure and verb forms. This writing is best defined as a collection of discrete sentences and/or questions loosely strung together. There is little evidence of deliberate organization. Intermediate Mid writers can be understood readily by natives used to the writing of non-natives. When Intermediate Mid writers attempt Advanced-level writing tasks, the quality and/or quantity of their writing declines and the message may be unclear.

INTERMEDIATE LOW

Writers at the Intermediate Low sublevel are able to meet some limited practical writing needs. They can create statements and formulate questions based on familiar material. Most sentences are recombinations of learned vocabulary and structures. These are short and simple conversational-style sentences with basic word order. They are written almost exclusively in present time. Writing tends to consist of a few simple sentences, often with repetitive structure. Topics are tied to highly predictable content areas and personal information. Vocabulary is adequate to express elementary needs. There may be basic errors in grammar, word choice, punctuation, spelling, and in the formation and use of non-alphabetic symbols. Their writing is understood by natives used to the writing of non-natives, although additional effort may be required. When Intermediate Low writers attempt to perform writing tasks at the Advanced level, their writing will deteriorate significantly and their message may be left incomplete.

The descriptions imply a pattern of development common to all language learners. They assume that a particular level of grammatical ability will always be associated with a particular level of lexical ability. This is, to say the least, highly questionable, and the scales have been criticised for not being based on research into the acquisition order of the various elements. Where scales are to be used to measure achievement, this criticism is, we believe, justified. If the different levels are not closely based on research into changes in performance over time, then their use is unlikely to lead to valid measures of achievement.

This is not to say that all scales need to be based on what is known of the way languages are learned. The ILR (Interagency Language Roundtable) Levels are similar in many ways to the ACTFL scales. The difference is that the ILR Levels were designed to assign individuals to a Level in order to determine whether their foreign language ability was sufficient for a particular job. The purpose is purely to measure proficiency, regardless of how it has been achieved. The ILR Levels (for speaking) are illustrated in the next chapter.

An issue which arises when using scales of the ACTFL (and ILR) kind is how to rate someone whose language is described partly by one level and partly by another (or others). What we decide must depend in part on the purpose of the assessment. If we are trying to find out if a person has sufficient language ability for, say, a diplomatic post, we might decide that we have to place them at the lowest level that (partly) describes their language. If the purpose is to measure achievement, we may be more willing to allow strengths in one area to compensate for weaknesses in another.

Analytic scoring

Methods of scoring which require a separate score for each of a number of aspects of a task are said to be *analytic*. The following scale, devised by John Anderson, is based on an oral ability scale found in Harris (1968).

GRAMMAR

6. Few (if any) noticeable errors of grammar or word order.
5. Some errors of grammar or word order which do not, however, interfere with comprehension.
4. Errors of grammar or word order fairly frequent; occasional re-reading necessary for full comprehension.
3. Errors of grammar or word order frequent; efforts of interpretation sometimes required on reader's part.
2. Errors of grammar or word order very frequent; reader often has to rely on own interpretation.
1. Errors of grammar or word order so severe as to make comprehension virtually impossible.

VOCABULARY

6. Use of vocabulary and idiom rarely (if at all) distinguishable from that of educated native writer.
5. Occasionally uses inappropriate terms or relies on circumlocutions; expression of ideas hardly impaired.
4. Uses wrong or inappropriate words fairly frequently; expression of ideas may be limited because of inadequate vocabulary.
3. Limited vocabulary and frequent errors clearly hinder expression of ideas.
2. Vocabulary so limited and so frequently misused that reader must often rely on own interpretation.
1. Vocabulary limitations so extreme as to make comprehension virtually impossible.

MECHANICS

6. Few (if any) noticeable lapses in punctuation or spelling.
5. Occasional lapses in punctuation or spelling which do not, however, interfere with comprehension.
4. Errors in punctuation or spelling fairly frequent; occasional re-reading necessary for full comprehension.
3. Frequent errors in spelling or punctuation; lead sometimes to obscurity.
2. Errors in spelling or punctuation so frequent that reader must often rely on own interpretation.
1. Errors in spelling or punctuation so severe as to make comprehension virtually impossible.

FLUENCY (STYLE AND EASE OF COMMUNICATION)

6. Choice of structures and vocabulary consistently appropriate; like that of educated native writer.
5. Occasional lack of consistency in choice of structures and vocabulary which does not, however, impair overall ease of communication.
4. 'Patchy', with some structures or vocabulary items noticeably inappropriate to general style.
3. Structures or vocabulary items sometimes not only inappropriate but also misused; little sense of ease of communication.
2. Communication often impaired by completely inappropriate or misused structures or vocabulary items.
1. A 'hotch-potch' of half-learned misused structures and vocabulary items rendering communication almost impossible.

FORM (ORGANISATION)

6. Highly organised; clear progression of ideas well linked; like educated native writer.
5. Material well organised; links could occasionally be clearer but communication not impaired.

4. Some lack of organisation; re-reading required for clarification of ideas.
3. Little or no attempt at connectivity, though reader can deduce some organisation.
2. Individual ideas may be clear, but very difficult to deduce connection between them.
1. Lack of organisation so severe that communication is seriously impaired.

SCORE:

Gramm: ____ + Voc: ____ + Mech: ____ + Fluency: ____ + Form: ____ = ____
(TOTAL)

There are a number of advantages to analytic scoring. First, it disposes of the problem of uneven development of sub-skills in individuals. Secondly, scorers are compelled to consider aspects of performance which they might otherwise ignore. And thirdly, the very fact that the scorer has to give a number of scores will tend to make the scoring more reliable. While it is doubtful that scorers can judge each of the aspects independently of the others (there is what is called a 'halo effect'), the mere fact of having (in this case) five 'shots' at assessing the student's performance should lead to greater reliability.

In Anderson's scheme, each of the components is given equal weight. In other schemes (such as that of Jacobs et al. (1981), below), the relative importance of the different aspects, as perceived by the tester (with or without statistical support), is reflected in weightings attached to the various components. Grammatical accuracy, for example, might be given greater weight than accuracy of spelling. A candidate's total score is the sum of the weighted scores.

The main disadvantage of the analytic method is the time that it takes. Even with practice, scoring will take longer than with the holistic method. Particular circumstances will determine whether the analytic method or the holistic method will be the more economical way of obtaining the required level of scorer reliability.

A second disadvantage is that concentration on the different aspects may divert attention from the overall effect of the piece of writing. Inasmuch as the whole is often greater than the sum of its parts, a composite score may be very reliable but not valid. Indeed the aspects that are scored separately (the 'parts'), presumably based on the theory of linguistic performance that most appeals to the author of any particular analytic framework, may not in fact represent the complete, 'correct' set of such aspects. To guard against this, an additional, impressionistic score on each composition is sometimes required of scorers, with significant discrepancies between this and the analytic total being investigated.

ESL COMPOSITION PROFILE				
STUDENT		DATE	TOPIC	
SCORE	LEVEL	CRITERIA		COMMENTS
CONTENT	30-27	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic		
	26-22	GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail		
	21-17	FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic		
	16-13	VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate		
ORGANIZATION	20-18	EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/ supported • succinct • well-organized • logical sequencing • cohesive		
	17-14	GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing		
	13-10	FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development		
	9-7	VERY POOR: does not communicate • no organization • OR not enough to evaluate		
VOCABULARY	20-18	EXCELLENT TO VERY GOOD: sophisticated range • effective word/ idiom choice and usage • word form mastery • appropriate register		
	17-14	GOOD TO AVERAGE: adequate range • occasional errors of word/ idiom form, choice, usage but meaning not obscured		
	13-10	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • meaning confused or obscured		
	9-7	VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate		
LANGUAGE USE	25-22	EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions		
	21-18	GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions but meaning seldom obscured		
	17-11	FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • meaning confused or obscured		
	10-5	VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate		
MECHANICS	5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing		
	4	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing but meaning not obscured		
	3	FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • meaning confused or obscured		
	2	VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate		
TOTAL SCORE		READER	COMMENTS	

It is worth noting a potential problem in Anderson's scale. This arises from the conjunction of frequency of error and the effect of errors on communication. It is not necessarily the case that the two are highly correlated. A small number of grammatical errors of one kind could have a much more serious effect on communication than a large number of another kind. This problem is not restricted to analytic scales, of course;

it is just as difficult an issue in more holistic scales. Research in the area of error analysis, particularly the study of error gravity, offers insights to those wishing to pursue the matter further.

An analytic scale widely used at college level in North America is that of Jacobs et al. (1981), reproduced on page 104. As can be seen, it has five components, 'content' being given the greatest weight and 'mechanics' the least. The weightings reflect the perceived importance of the different components in writing at college level. They would not necessarily be appropriate for testing the writing at a more elementary level, where control of mechanics might be considered more important. Note also that, except in the case of mechanics, a range of scores is associated with each descriptor, allowing the scorer to vary the score assigned in accordance with how well the performance fits the descriptor.

The choice between holistic and analytic scoring depends in part on the purpose of the testing. If diagnostic information is required directly from the ratings given, then analytic scoring is essential².

The choice also depends on the circumstances of scoring. If it is being carried out by a small, well-knit group at a single site, then holistic scoring, which is likely to be more economical of time, may be the most appropriate. But if scoring is being conducted by a heterogeneous, possibly less well trained group, or in a number of different places, analytic scoring is probably called for. Whichever is used, if high accuracy is sought, multiple scoring is desirable.



STEPS IN CONSTRUCTING A RATING SCALE

Constructing a valid rating scale is no easy matter. What follows is a practical guide to scale construction, assuming that it will not be possible to carry out extensive empirical research or use advanced statistical methods. It can also be used for the construction of oral rating scales.

1. Ask: What is the purpose of the testing?
 - How many distinctions in ability have to be made?
 - How will the 'scores' be reported?
 - What are the components of the ability which you want to measure?
 - Is it intended to provide feedback? If so, how detailed must it be?
2. In the light of the answers to the previous questions, decide:
 - whether scoring should be analytic or holistic, or both;
 - how many components the scale should have;
 - how many separate levels the scale should have.
3. Search for existing scales that are similar to what you need or that can contribute to the construction of your scale.
4. Modify existing scales to suit your purpose.
5. Trial the scale you have constructed and make whatever modifications prove necessary. If possible, retri al the scale before calibrating it.

² Where there is holistic scoring, a checklist may be used for raters to indicate particular strengths and weaknesses (see the box on page 99).

Any scale which is used, whether holistic or analytic, should reflect the particular purpose of the test and the form that the reported scores on it will take. Because valid scales are not easy to construct, it is eminently reasonable to begin by reviewing existing scales and choosing those that are closest to one's needs. It should go without saying, however, that the chosen scales will almost certainly need to be adapted for the situation in which they are to be used.

Finally in this section, it is also worth pointing out that since scales are in effect telling candidates 'These are the criteria by which we will judge you', their potential for backwash is considerable, provided that candidates are made aware of them.

Calibrate the scale to be used

Any scale which is to be used should first be calibrated. As said in the previous chapter, this means collecting samples of performance collected under test conditions, and covering the full range of the scales. Members of the testing team (or another set of experts) then look at these samples and assign each of them to a point (or points in the case of an analytic scale) on the relevant scale. The assigned samples provide reference points for all future uses of the scale, as well as being essential training materials.

Select and train scorers

Not everyone is equally good at rating written work, even with training. Trainee scorers should be expert users of the language being tested. They should be sensitive to language, have had experience of teaching writing and marking written work. It is also helpful if they have had training in testing.

We would recommend that training be carried out in three stages, each to be held on a separate day, though we recognise that this is often impractical. If possible, the training should take place on three consecutive days. A possible outline for training follows.



OUTLINE FOR TRAINING

Training Stage 1 Background and Overview

- Background and rationale.
- Trainees are given a copy of the writing handbook and taken through its contents.
- Examples of writing are given, one at a time, with one at each level. Participants compare relevant descriptors with the pieces of work. There is discussion about each piece of work and how it should be rated. The trainer will have an agreed completed rating sheet for each piece of work.
- All pieces of work should be on the same topic, for all stages of the training.

- There should be at least one case where quite different pieces of work are assigned to the same level. For example, one may be strong in grammar and vocabulary but not very well organised, while the other is well structured and coherent but contains significantly more grammatical errors.
- Trainees are asked to study the handbook and the sample compositions before the second stage of training.

Training Stage 2

- Queries arising from the handbook are answered.
- A set of calibrated pieces of work is given to each trainee. (All levels should be covered, with extra examples being in the middle of the range.) Trainees are asked to complete a rating sheet independently, assigning each piece of work to a level.
- A discussion follows the assignment of all pieces of work to levels.
- The trainer has an agreed completed rating sheet for each piece of work. This cannot be challenged.
- All completed rating sheets are kept as a record of the trainees' performance.

Training Stage 3 Assessment

- As stage 2, except that there is no discussion.
- An agreed level of accuracy is required for someone to become a rater. Those who do not achieve it do not become raters.

Automated scoring

The use of computers to score writing is controversial, particularly in high-stakes testing. However, although automated scoring is currently used mostly by large-scale testing organisations, it is likely to reach a wider audience in the future. With this in mind, there follows a brief summary of the issues involved with automated scoring of writing.

The advantages of automated scoring are perhaps obvious. Computers can score several pieces of writing much faster than a human can. After a certain length of time, the use of computers will also work out cheaper than employing human markers. In addition to time and cost savings, when presented with a piece of writing, an automated scoring system is guaranteed to assign the same grade today as it did last week and the week before that. In this sense, a computer should be able to achieve perfect reliability.

There are however serious validity concerns about allowing assessment of writing to be done by computers³. Perhaps the most significant drawback

³ The embracing of automated scoring is reminiscent of the predilection of some testers in the past for multiple choice items, which similarly valued reliability and economy over validity.

is that computers are currently unable to rate the higher-level features of writing such as argument development or the logical connection between ideas. Connected with this is readability, and it is questionable whether a computer is able to effectively assess such a human quality. Furthermore, due to the nature of computer algorithms, test-takers can potentially trick automated scoring systems by including language they know will be scored highly, while not necessarily providing appropriate content. There are also concerns about the potential negative backwash due to teachers and students focusing on the aspects of writing they believe will gain high grades from an automated system. Finally, there is a more general concern about handing over control to computers, especially when the precise ways in which they assess writing are not clear or understandable to many of us. See Further reading for more information.

With computers currently able to perform some very useful functions but unable to totally assess writing satisfactorily, we believe that when automated scoring is used, it should be complemented by human raters. This is currently the case with *TOEFL*[®], where human raters focus on content and meaning, with automated scoring focusing on linguistic features. Similarly in the classroom, there is great potential for automated scoring to focus on the simpler aspects of writing, thereby freeing the teacher to focus on higher-level aspects. This should be particularly useful in formative assessment.

Follow acceptable scoring procedures

It is assumed that scorers have already been trained. Once the test is completed, a search should be made to identify 'benchmark' scripts that typify key levels of ability on each writing task (in the case of the English-medium university referred to above, these were 'adequate' and 'not adequate'; another test might require examples at all levels)⁴. Copies of these should then be presented to the scorers for an initial scoring. Only when there is agreement on these benchmark scripts should scoring begin. Each task of each student should be scored independently by two or more scorers (as many scorers as possible should be involved in the assessment of each student's work), the scores being recorded on separate sheets. A third, senior member of the team should collate scores and identify discrepancies in scores awarded to the same piece of writing. Where these are small, the two scores can be averaged; where they are larger, senior members of the team will decide the score. It is also worth looking for

⁴ Interestingly, we have noticed that when markers are scoring a large number of written compositions according to a set of criteria, they will sometimes begin by ordering the compositions from high to low. This is understandable, as it can help a marker to benchmark. It should be discouraged, however, since it is essentially turning what should be a criterion-referenced test into a norm-referenced test.

large discrepancies between an individual's performance on different tasks. These may accurately reflect their performance, but they may also be the result of inaccurate scoring.

It is important that scoring should take place in a quiet, well-lit environment. Scorers should not be allowed to become too tired. While holistic scoring can be very rapid, it is nevertheless extremely demanding if concentration is maintained.

Multiple scoring should ensure scorer reliability, even if not all scorers are using quite the same standard. Nevertheless, once scoring is completed, it is useful to carry out simple statistical analyses to discover if anyone's scoring is unacceptably aberrant. One might find, for example, that one person is rating higher (or lower) than the others. This can be brought to their attention. If someone's rating is markedly wayward, but not in one direction, it may be wise not to ask them to rate work in future.

Comparative Judgement

As mentioned in Chapter 3, Comparative Judgement (CJ) is an approach to the marking of written work which has gained some currency in recent years. It involves groups of judges working individually, each judge being given two randomly chosen scripts at a time (on paper or on computer) and being asked simply to decide which of the two is better. Those scripts which are deemed to be better are termed 'winners'; those that are not judged better are 'losers'. When all of the scripts have been judged in this way, the process is repeated, with the difference that winners are compared with winners, and losers compared with losers. After four iterations of this process, it is possible to assign all of the scripts to a single scale.

The advantages of CJ are said to be high inter-scorer reliability, practicality and, because the setting of tasks is not restricted by the kind of considerations we have identified above, potentially high content validity. Drawbacks are that scores arrived at in this way are not criterion related and the procedure is not capable of giving diagnostic information in the form of feedback.

Feedback

There will be many situations in which feedback to the candidates on their performance will be useful. The provisional content of a feedback pro forma can be decided during calibration. Here, for example, is a list of the elements that were thought worthy of inclusion at calibration sessions which one of us attended.

In addition to feedback on linguistic features (e.g. grammar; vocabulary, limited or used inappropriately), the following elements should be included on the feedback pro forma:

Non-writing-specific:

- incomplete performance of the task in terms of:
 1. topic: not all parts addressed very superficial treatment
 2. operations called for (e.g. compare and contrast)
- pointless repetition

Writing-specific:

- misuse of quotation marks
- inappropriate underlining
- capitalization
- style conventions
- failure to split overlong sentences
- inappropriate use of sentence fragments
- handwriting

Computer-based feedback

The use of computers in giving feedback is much less controversial than automated scoring. There are several benefits, which generally apply to formative, rather than summative assessment. Instead of waiting for a teacher to read a piece of writing, correct errors and choose areas to focus on, students can receive automated feedback instantly. It is also possible for students to submit their work numerous times, correcting and improving it on each submission. Another advantage is the potential for anonymity which means students are more likely to take risks in their writing. This willingness among students to risk making mistakes is not always apparent when they are expecting personalised feedback from their teacher.

There are a number of websites which allow users to submit a piece of writing and receive immediate feedback. One of the more well-known is Cambridge English *Write & Improve*. With this free online service, users can choose from a selection of writing tasks aimed at students at varying levels of proficiency. Once they have submitted their response, their writing is assigned a *CEFR* level from A1 to C2. In addition, sections of writing identified as 'problematic' are highlighted. Students are encouraged to rework these sections and resubmit.

While the benefits of automated feedback programs like *Write & Improve* are outlined above, its limitations become immediately apparent. When feedback is as general as having a word or sentence highlighted as problematic, with no detail as to *why*, it can leave students confused. Whereas in a student-teacher dynamic this confusion can lead to useful conversations where problems are analysed and alternatives are elicited, none of this is possible with automated feedback tools. This lack of dialogue, an important feature of feedback, is a frustrating aspect of using such a program. Arguably, this perfectly illustrates how such technology can be useful as a teaching aid but is currently far from being able to recreate or replace the work a human can do.



READER ACTIVITIES

1. Following the advice given in this chapter, construct two writing tasks appropriate to a group of students with whom you are familiar. Carry out the tasks yourself. If possible, get the students to do them as well. Do any of the students produce writing different in any significant way from what you hoped to elicit? If so, can you see why? Would you wish to change the tasks in any way?
2. Visit the free online written feedback service, Cambridge English *Write & Improve* and follow the instructions for submitting a piece of writing. What is your opinion of the feedback you receive? Is it accurate? Is it useful? Would you consider using this service with your students?
3. Think of a time when you were trained as a rater (if you ever were). How similar was the training to the outline presented on pages 106–107? If there were differences, why do you think that was?
4. This activity is best carried out with colleagues. Score the following three short compositions on how to increase tourism, using each of the scales presented in the chapter. Which do you find easiest to use, and why? How closely do you and your colleagues agree on the scores you assign? Can you explain any large differences? Do the different scales place the compositions in the same order? If not, can you see why not? Which of the scales would you recommend in what circumstances?

1. Nowadays a lot of countries tend to develop their tourism's incomes, and therefore tourism called the factory without chimney. Turkey, which undoubtedly needs foreign money, tries to increase the number of foreign tourists coming to Turkey. What are likely to do in order to increase this number.

At first, much more and better advertising should do in foreign countries and the information offices should open to inform the people to decide to come Turkey. Secondly, improve facilities, which are hotels, transportation and communication. Increase the number of hotels, similarly the number of public transportation which, improve the lines of communication. Thirdly which is important as two others is training of personnel. This is also a basic need of tourism, because the tourist will want to see in front of him a skilled guides or a skilled hotel managers. The new school will open in order to train skilled personnel and as well as theoretic knowledges, practice must be given them.

The countries which are made available these three basic need for tourists have already improved their tourism's incomes. Spain is a case in point or Greece. Although Turkey needs this income; it didn't do any real attempts to achieve it. In fact all of them should have already been done, till today. However it is late, it can be begin without losing any time.

2. *A nation can't make improvements, if it doesn't let the minds of their people breathe and expand to understand more about life than what is at the end of the street, this improvement can be made by means of tourism.*

There are several ways to attract more people to our country. First of all, advertisements and information take an important place. These advertisements and information should be based on the qualities of that place without exaggeration. The more time passes and the more information tourists gather about one country, the more assured they can be that it will be a good experience. People travel one place to another in order to spend their holiday, to see different cultures or to attend conferences. All of these necessitate facilities. It is important to make some points clear. Hotel, transportation and communication facilities are a case in point. To some extent, we can minimize the difficulties by means of money. Furthermore, this situation does not only depend on the financial situation, but also behaviors towards the tourists. Especially, a developing country should keep in mind the challenge of the future rather than the mistakes of the past, in order to achieve this, the ways of training of personnel may be found. The most important problem faced by many of countries is whether the decisions that must be made are within the capabilities of their education system. Educating guides and hotel managers are becoming more and more important.

As a result, it should once more be said that, we may increase the number of foreign tourists coming to Turkey by taking some measures. Advertisement, information, improving facilities and training personnel may be effective, but also all people should be encouraged to contribute this event.

3. *Tourism is now becoming a major industry throughout the world. For many countries their tourist trade is an essential source of their revenue.*

All countries have their own particular attractions for tourists and this must be kept in mind when advertising Turkey abroad. For example Turkey, which wants to increase the number of foreign tourists coming must advertise its culture and sunshine.

Improving facilities like hotels, transportation and communication play an important role on this matter more. Hotels can be built and available ones can be kept clean and tidy.

New and modern transportation systems must be given to foreign tourists and one more, the communication system must work regularly to please these people.

Tourists don't want to be led around like sheep. They want to explore for themselves and avoid the places which are packed out with many other tourists. Because of that there must be their trained guides on their tours through anywhere and on the other hand hotel managers must be well trained. They must keep being kind to foreign tourists and must know English as well.

If we make tourists feel comfortable in these facts, tourism will increase and we will benefit from it.

5. This activity is also best carried out with colleagues. Construct a holistic writing scale and an analytic writing scale appropriate for use with the group of students you have already identified. If possible, score the students' efforts on the two tasks (Activity 1), using both methods. Look at differences between scorers and between methods, as in the previous activity. What changes would you make in the scales? Which of the two scales would be most useful for your purposes?

(Hughes et al. 1987)

FURTHER READING

General

Weigle (2002) is a thorough treatment of the assessment of writing. It includes chapters on portfolio assessment and on the future of writing assessment (including the use of computers as raters). Jacobs et al. (1981) (available online as a pdf), from which one of the scales presented in this chapter was taken, is also recommended. For an overview of the practical issues involved in the assessment of writing, see Coombe (2010). Godshalk et al. (1966) describes in detail the development of an indirect test of writing ability.

Analysis of tests

Shaw and Weir (2007) provide an explanation of the Cambridge approach to writing assessment with reference to the Cambridge suite of tests. Similarly, Chapelle et al. (2008) describes and analyses a major revision of the TOEFL® test.

Scales and scoring

North and Schneider (1998) report on the development of a language proficiency scale. Council of Europe (2001) contains a number of scales (not only of writing ability) which are potentially useful to test constructors needing to create their own, as well as an annotated bibliography on language proficiency scaling. To see how experienced raters assign varying value to different criteria, see Eckes (2008). Johnson and Hamp-Lyons (1995) point to problems with holistic scoring. Bouwer et al. (2014) shows how an increase in the number of texts and genres in writing tasks gives us a more accurate impression of writing proficiency. Jennings et al. (1999) found that allowing a choice of topic did not make a difference to test-takers' scores (but one should be wary about extrapolating from one study in one situation). Elder et al. (2007) examine the effectiveness of online rater training programmes. Weigle (1994) reports the effects of training on raters of ESL compositions. Pollitt (2012) provides a good, clear introduction to Adaptive Comparative Judgement (a version of Comparative Judgement).

Automated scoring

See Weigle (2013) for an explanation of how some of the most common automated essay grading systems work, a description of what they can and cannot do, as well as a summary of their place in the assessment of writing. However, readers should bear in mind the ever-changing nature of technological advances. The ETS (Educational Testing Services) website describes the capabilities of the e-rater tool and provides an extensive list of research papers into automated writing evaluation. Pearson Assessments have published reports online about their automated scoring systems, where they also give details of a publicly available version.

Feedback

Hyland and Hyland (2006) covers a wide range of issues involved with giving feedback on written work.

For a focus on error treatment in student writing, see Ferris (2002).

10

Testing speaking

The assumption is made in this chapter that the objective of teaching spoken language is the development of the ability to interact successfully in that language, and that this involves comprehension as well as production. It is also assumed that at the earliest stages of learning formal testing of this ability will not be called for, informal observation providing any diagnostic information that is needed.

The basic problem in testing speaking ability is essentially the same as for testing writing.

1. We want to set tasks that form a representative sample of the population of speaking tasks that we expect candidates to be able to perform.
2. The tasks should elicit behaviour which truly represents the candidates' ability.
3. The samples of behaviour can and will be scored validly and reliably.

Following the pattern of the previous chapter, we shall deal with each of these in turn.

Representative tasks

Specify all possible content

We will begin by looking at the specified content of the *Cambridge English B2 First*.

Functions: express opinions, justify opinions, speculate, summarise, reaching a decision through negotiation, invite opinions and ideas, discuss, evaluate, comparing, describing, exchanging ideas, agreeing and disagreeing, suggesting

Types of text: interview, collaborative task, discussion, individual 'long turn'

Addressees: interlocutor (an examiner) and a fellow candidate

Topics: personal information (e.g. work, leisure time, future plans)

Dialect, accent and style: not specified.¹

¹ These specifications are derived from the *Cambridge English B2 First Handbook*, where they are occasionally referred to using different terms.

These content specifications may be compared with those for a test with which we have been concerned. The categorisation of the operations (here referred to as skills) is based on Bygate (1987).

SKILLS

Informational skills

Candidates should be able to:

- provide personal information
- provide non-personal information
- describe sequence of events (narrate)
- give instructions
- make comparisons
- give explanations
- present an argument
- provide required information
- express need
- express requirements
- elicit help
- seek permission
- apologise
- elaborate an idea
- express opinions
- justify opinions
- complain
- speculate
- analyse
- make excuses
- paraphrase
- summarise (what they have said)
- make suggestions
- express preferences
- draw conclusions
- make comments
- indicate attitude

Interactional skills

Candidates should be able to:

- express purpose
- recognise other speakers' purpose
- express agreement
- express disagreement
- elicit opinions
- elicit information
- question assertions made by other speakers
- modify statements or comments

- justify or support statements or opinions of other speakers
- attempt to persuade others
- repair breakdowns in interaction
- check that they understand or have been understood correctly
- establish common ground
- elicit clarification
- respond to requests for clarification
- correct themselves or others
- indicate understanding (or failure to understand)
- indicate uncertainty

Skills in managing interactions

Candidates should be able to:

- initiate interactions
- change the topic of an interaction
- share the responsibility for the development of an interaction
- take their turn in an interaction
- give turns to other speakers
- come to a decision
- end an interaction

Types of text

- Presentation (monologue)
- Discussion
- Conversation
- Service encounter
- Interview

Other speakers (addressees)

- may be of equal or higher status
- may be known or unknown

Topics Topics which are familiar and interesting to the candidates

Dialect Standard British English or Standard American English

Accent RP, Standard American

Style Formal and informal

Vocabulary range Non-technical except as the result of preparation for a presentation

Rate of speech Will vary according to task

It can be seen that this second set of content specifications is rather fuller than the first. What is more, splitting the skills into three categories (informational, interactional and management), as it does, should help in creating tasks which will elicit a representative sample of each. In our

view, the greater the detail in the specification of content, the more valid the test is likely to be. Readers may wish to select elements from the two sets of specifications for their own purposes.

Include a representative sample of the specified content when setting tasks

Any one speaking test should sample from the full specified range. The reasons for doing this are the same as those given in the previous chapter. Let us look at the materials for a recent *Cambridge English B2 First* test.

Part 1

2 minutes (3 minutes for groups of three)

Good morning/afternoon/evening. My name is and this is my colleague

And your names are?

Can I have your mark sheets, please?

Thank you.

- Where are you from, (*Candidate A*)?
- And you, (*Candidate B*)?

First we'd like to know something about you.

Select one or more questions from any of the following categories, as appropriate.

Likes and dislikes

- How do you like to spend your evenings? (What do you do?) (Why?)
- Do you prefer to spend time on your own or with other people? (Why?)
- Tell us about a film you really like.
- Do you like cooking? (What sort of things do you cook?)

Special occasions

- Do you normally celebrate special occasions with friends or family? (Why?)
- Tell us about a festival or celebration in (*candidate's country*).
- What did you do on your last birthday?
- Are you going to do anything special this weekend? (Where are you going to go?) (What are you going to do?)

Media

- How much TV do you watch in a week? (Would you prefer to watch more TV than that or less?) (Why?)
- Tell us about a TV programme you've seen recently.
- Do you use the internet much? (Why? / Why not?)
- Do you ever listen to the radio? (What programmes do you like?) (Why?)

1 Helping others
2 Gardens

Part 2

4 minutes (6 minutes for groups of three)

Interlocutor

In this part of the test, I'm going to give each of you two photographs. I'd like you to talk about your photographs on your own for about a minute, and also to answer a question about your partner's photographs.

(Candidate A), it's your turn first. Here are your photographs. They show **people who are helping other people in different situations**.

Place **Part 2** booklet, open at **Task 1**, in front of Candidate A.

I'd like you to compare the photographs, and say **how important it is to help people in these situations**.

All right?

Candidate A

🕒 1 minute

Interlocutor

Thank you.

(Candidate B), **do you find it easy to ask for help when you have a problem?**
(Why? / Why not?)

Candidate B

🕒 approximately 30 seconds

Interlocutor

Thank you. (Can I have the booklet, please?) Retrieve **Part 2** booklet.

Now, (Candidate B), here are your photographs. They show **people spending time in different gardens**.

Place **Part 2** booklet, open at **Task 2**, in front of Candidate B.

I'd like you to compare the photographs, and say **what you think the people are enjoying about spending time in these gardens**.

All right?

Candidate B

🕒 1 minute

Interlocutor

Thank you.

(Candidate A), **which garden would you prefer to spend time in?** (Why?)

Candidate A

🕒 approximately 30 seconds

Interlocutor

Thank you. (Can I have the booklet, please?) Retrieve **Part 2** booklet.

How important is it to help people in these situations?

1



What are the people enjoying about spending time in these gardens?

2



21 Holiday resort**Part 3** 4 minutes (5 minutes for groups of three)**Part 4** 4 minutes (6 minutes for groups of three)**Part 3**

Interlocutor Now, I'd like you to talk about something together for about two minutes. *(3 minutes for groups of three).*

I'd like you to imagine that a town wants more tourists to visit. Here are some ideas they're thinking about and a question for you to discuss. First you have some time to look at the task.

Place Part 3 booklet, open at Task 21, in front of the candidates. Allow 15 seconds.

Now, talk to each other about **why these ideas would attract more tourists to the town.**

Candidates

⌚ 2 minutes
(3 minutes for
groups of three)

.....

Interlocutor Thank you. Now you have about a minute to decide **which idea would be best for the town.**

Candidates

⌚ 1 minute
(for pairs and
groups of three)

.....

Interlocutor Thank you. (Can I have the booklet, please?) *Retrieve Part 3 booklet.*

Part 4

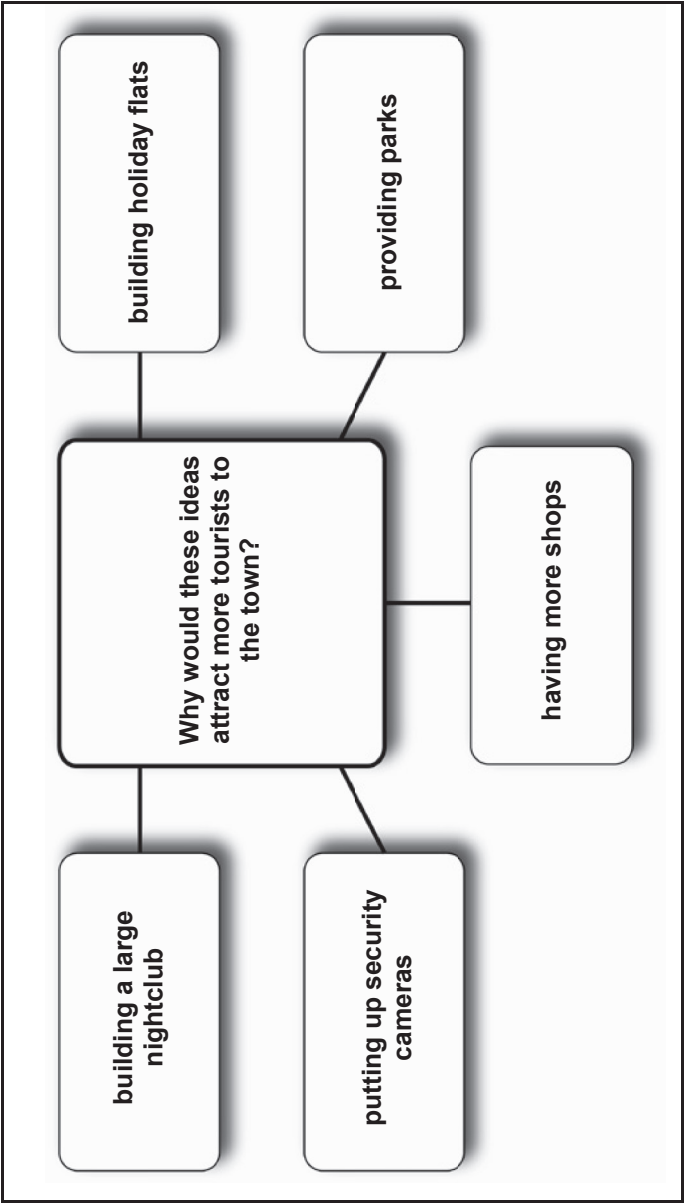
Interlocutor *Use the following questions, in order, as appropriate:*

- **Do you think you have to spend a lot of money to have a good holiday?** (Why? / Why not?)
- **Some people say we travel too much these days and shouldn't go on so many holidays. What do you think?**
- **Do you think people have enough time for holidays these days?** (Why? / Why not?)
- **Why do you think people like to go away on holiday?**
- **What do you think is the biggest advantage of living in a place where there are a lot of tourists?**
- **What can people do to have a good holiday in (candidate's country)?** (Why?)

Select any of the following prompts, as appropriate:

- **What do you think?**
- **Do you agree?**
- **And you?**

Thank you. That is the end of the test.



It is interesting to try to predict which of the functions listed in the specifications would be elicited by these tasks. You might want to attempt to do this before reading any further. Looking at them ourselves, we thought that in performing the tasks, the speakers were quite likely to express and justify opinions, speculate, describe, discuss, compare, suggest, exchange ideas, agree and disagree, and negotiate. You may notice that we have listed almost all the functions from the test specifications, suggesting these tasks do a good job of including a representative sample.

It is also interesting to notice the scripted nature of what the interlocutor says during the interview. A tight script like this is designed to improve reliability and fairness since each candidate is being given the same input and opportunity to perform. Interlocutor scripts also make it easier to control the content of the task and, providing the prompts are well designed, it should be easier to ensure content validity. However, these benefits depend on the prompts eliciting what they are intended to elicit. During the test development, the prompts should be checked and trialled and, if necessary, altered.

The drawback to such a tightly controlled task is that the lack of flexibility may prevent candidates from performing to their maximum potential. A balance should be found so that every candidate is given an equal opportunity to show what they can do. Much may depend on the interlocutor; ideally, they should be aware of why they are asking each question, and which functions they are likely to elicit.

Elicit a valid sample of speaking ability

Choose appropriate techniques

Three general formats are presented here: interview; interaction with fellow candidates; responses to audio- or video-recorded stimuli.

Format 1 Interview

Perhaps the most common format for the testing of oral interaction is the interview. In its traditional form, however, it has at least one potentially serious drawback. The relationship between the tester and the candidate is usually such that the candidate speaks as to a superior and is unwilling to take the initiative, although an interlocutor may take steps to minimise this by creating a less formal atmosphere and establishing a more equal power dynamic. As a result of the typically imbalanced nature of the relationship, only one style of speech is elicited, and many functions (such as asking for information) are not represented in the candidate's performance. It is possible, however, to get round this problem by introducing a variety of elicitation techniques into the interview situation.

Useful techniques are:

Questions and requests for information

Yes/No questions should generally be avoided, except perhaps at the very beginning of the interview, while the candidate is still warming up. Performance of various operations (of the kind listed in the two sets of specifications above) can be elicited through requests of the kind:

Can you explain to me how/why ...? and

Can you tell me what you think of ...?

Requests for elaboration

For example:

What exactly do you mean?, Can you explain that in a little more detail?, What would be a good example of that? Tell me more.

Appearing not to understand

This is most appropriate where the interviewer really isn't sure of what the candidate means but can also be used simply in order to see if the candidate can cope with being misunderstood. The interviewer may say, for example, *I'm sorry, but I don't quite follow you.*

Invitation to ask questions

Is there anything you'd like to ask me?

Interruption

To see how the candidate deals with this.

Abrupt change of topic

To see how the candidate deals with this.

Pictures

Single pictures are particularly useful for eliciting descriptions. Series of pictures (or video sequences) form a natural basis for narration (the series of pictures on pages 94–95 for example). Pictures can also be used as starters for a discussion, as seen in the *B2 First* example earlier in this chapter.

Role play

Candidates can be asked to assume a role in a particular situation. This allows the ready elicitation of other language functions. There can be a series of brief items, such as:

A friend invites you to a party on an evening when you want to go to a gym class. Thank the friend (played by the tester) and refuse politely.

Or there can be a more protracted exchange:

You want your mother (played by the tester) to increase your pocket money. She is resistant to the idea. Try to make her change her mind.

You want to fly from London to Paris on 13 March, returning a week later. Get all the information that you need in order to choose your flights from the travel agent (played by the tester).

In our experience, however, where the aim is to elicit 'natural' language and an attempt has been made to get the candidates to forget, to some extent at least, that they are being tested, role play can destroy this illusion. We have found that some candidates, rather than responding to the situation as if it were one they were actually facing, will resort to uttering half-remembered snatches of exchanges once learned by rote. An added drawback is the tendency for less confident students to struggle with role-playing activities, which is likely to compromise the validity of the assessment.

Interpreting

It is not intended that candidates should be able to act as interpreters (unless that is specified). However, simple interpreting tasks can test both production and comprehension in a controlled way. If there are two testers, one of the testers acts as a monolingual speaker of the candidate's native language, the other as a monolingual speaker of the language being tested. Situations of the following kind can be set up:

The monolingual language speaker wants to invite a foreign visitor to his or her home for a meal. The candidate has to convey the invitation and act as an interpreter for the subsequent exchange.

Comprehension can be assessed when the candidate attempts to convey what the visitor is saying, and indeed unless some such device is used, it is difficult to obtain sufficient information on candidates' powers of comprehension. Production is tested when the candidate tries to convey the meaning of what the monolingual speaker says.

Prepared monologue

In the first edition of this book we said that we did not recommend prepared monologues as a means of assessing candidates' speaking ability. This was because we knew that the technique was frequently misused, often with candidates memorising the monologues. What we should have said is that it should only be used where the ability to make prepared presentations is something that the candidates will need, for example on certain university courses. Thus it could be appropriate in a proficiency test for teaching assistants, or in an achievement test where the ability to make presentations is an objective of the course.

Reading aloud

This is another technique the use of which we discouraged in the first edition, pointing out that there are significant differences amongst expert

speakers in the ability to read aloud, and that interference between the reading and the speaking skills was inevitable. But, if that ability is needed or its development has been a course objective, use of the technique may be justified. The use of reading aloud tasks is most beneficial when assessing pronunciation, and in our experience, more useful in assessing lower-level students, though it is currently used in one major test of English at university entry level.

Format 2 Interaction with fellow candidates

An advantage of having candidates interacting with each other is that it should elicit language that is appropriate to exchanges between equals, which may well be called for in the test specifications. It may also elicit better performance, inasmuch as the candidates may feel more confident than when dealing with a dominant, seemingly omniscient interviewer.

There is a problem, however. The performance of one candidate is likely to be affected by that of the others. For example, an assertive and insensitive candidate may dominate and not allow another candidate to show what he or she can do. If interaction with fellow candidates is to take place, the pairs should be carefully matched whenever possible. In general, we would advise against having more than two candidates interacting, as with larger numbers the chance of a diffident candidate failing to show their ability increases.

Possible techniques are:

Discussion

An obvious technique is to set a task which demands discussion between the two candidates, as in the Test of Oral Interaction above. Tasks may require the candidates to go beyond discussion and, for example, take a decision.

Role play

Role play can be carried out by two candidates with the tester as an observer. For some roles this may be more natural than if the tester were involved. It may, for example, be difficult to imagine the tester as 'a friend'. However, we believe that the doubts about role play expressed above still apply.

Format 3 Responses to audio or video recordings

Uniformity of elicitation procedures can be achieved through presenting all candidates with the same computer-generated or audio-/video-recorded stimuli (to which the candidates themselves respond into a microphone). This format, often described as 'semi-direct', ought to promote reliability. It can also be economical where a language laboratory is available, since large numbers of candidates can be tested at the same time. The obvious disadvantage of this format is its inflexibility: there is no way of following up candidates' responses.

There are a variety of techniques which can be used. These include:

Described situations

For example:

You are supposed to meet your friend outside a restaurant and they are already 45 minutes late. You decide to call them. What do you say?

Remarks in isolation to respond to

For example:

The candidate hears, 'I'm afraid I'm not able to come to your birthday party on Saturday. Sorry.'

or 'There are a couple of good films on at the cinema tonight.'

Simulated conversation

For example:

The candidate is given written information about a football match in Newcastle.

Newcastle United v Liverpool

Saturday 15 March

16.30

Tickets: £30 £50 £75

The candidate is given time to become familiar with the information, before being told that she or he wants to go to the match with a friend, Simon. Simon lives near to the Newcastle football ground but does not know about the match.

Simon's part in the conversation is played, and the candidate has to respond to what she or he hears.

The candidate hears:

Simon: Hello. What can I do for you?

PAUSE

Simon: That should be a good game. What day is it on?

PAUSE

Simon: And what time is it? Is it an afternoon or evening game?

PAUSE

Simon: OK, I'll get us two tickets. How much do you want to pay? How much are the cheapest?

PAUSE

Simon: Great. That's what I'll get. We don't need the best seats. I'm looking forward to it! I'll see you outside the ground.

PAUSE

Automated scoring

Computers are now used not only to provide audio prompts but, perhaps surprisingly, also to score the spoken performance of candidates. Not surprisingly, the range of responses that can be scored in this way is rather small. In Pearson's *Versant English Test*, for example, candidates are required to: read a number of sentences out loud; answer short questions such as *What is frozen water called?*; reorder phrases to form a correct sentence; briefly retell a story. Responses on these items are scored by computer algorithms. Significantly, however, candidates are then asked two questions that require longer responses (such as, *Do you like playing more in individual or in team sports?*), which are "not scored, but are available for review by authorized listeners"².

The advantages of computer scoring of speaking performance are obvious. It is practical (fast, economical) and potentially reliable. But for the present at least, in our opinion its validity is questionable, except when the purpose of testing is to obtain only a rough-and-ready estimate of speaking ability, such as for placement purposes.

Practical advice on conducting a speaking test



PLAN AND STRUCTURE THE TESTING CAREFULLY

1. Make the speaking test as long as is feasible. It is unlikely that much reliable information can be obtained in less than about 15 minutes, while 30 minutes can probably provide all the information necessary for most purposes. As part of a placement test, however, a five- or ten-minute interview should be sufficient to prevent gross errors in assigning students to classes.
2. Plan the test carefully. While one of the advantages of individual speaking testing is the way in which procedures can be adapted in response to a candidate's performance, the tester should nevertheless have some pattern to follow. It is a mistake to begin, for example, an interview with no more than a general idea of the course that it might take. Simple plans of the kind illustrated below can be made and consulted unobtrusively during the interview.

INTRO: Name, etc.

How did you get here today? traffic problems?

School: position, class sizes, children

Typical school day; school holidays

Three pieces of advice to new teachers

Examinations and tests

Tell me about typical errors in English

How do you teach ... present perfect v. past tense; future time reference; conditionals

What if... you hadn't become a teacher

... you were offered promotion

INTERPRETING: How do I get onto the internet?

How do I find out about the cheapest flights to Europe?

² This is from the *Versant* test description and validation summary.

NEWSPAPER: (look at the headlines)

EXPLAIN IDIOMS: For example, 'Once in a blue moon' or 'See the light'

3. Give the candidate as many 'fresh starts' as possible. This means a number of things. First, if possible and if appropriate, more than one format should be used. Secondly, again if possible, it is desirable for candidates to interact with more than one tester. Thirdly, within a format there should be as many separate 'items' as possible. Particularly if a candidate gets into difficulty, not too much time should be spent on one particular function or topic. At the same time, candidates should not be discouraged from making a second attempt to express what they want to say, possibly in different words.
4. Use a second tester for interviews. Because of the difficulty of conducting an interview and of keeping track of the candidate's performance, it is very helpful to have a second tester present. This person can not only give more attention to how the candidate is performing but can also elicit performance which they think is necessary in order to come to a reliable judgement. The interpretation task suggested earlier needs the co-operation of a second tester.
5. Set only tasks and topics that would be expected to cause candidates no difficulty in their own language. As teachers, many of us will have seen otherwise strong students struggle with tasks such as debates or presentations. This is often caused by non-linguistic issues such as a lack of confidence.
6. Carry out the interview in a quiet room with good acoustics.
7. Put candidates at their ease so that they can show what they are capable of. Individual speaking tests will always be particularly stressful for candidates. It is important to be pleasant and reassuring throughout, showing interest in what the candidate says through both verbal and non-verbal signals. It is especially important to make the initial stages of the test well within the capacities of all reasonable candidates. Interviews, for example, can begin with straightforward requests for personal (but not too personal) details, remarks about the weather, and so on. Testers should avoid constantly reminding candidates that they are being assessed. In particular they should not be seen to make notes on the candidates' performance during the interview or other activity. For the same reason, transitions between topics and between techniques should be made as natural as possible. The interview should be ended at a level at which the candidate clearly feels comfortable, thus leaving him or her with a sense of accomplishment.
8. Collect enough relevant information. If the purpose of the test is to determine whether a candidate can perform at a certain predetermined level, then, after an initial easy introduction, the test should be carried out at that level. If it becomes apparent that a candidate is clearly very weak and has no chance of reaching the criterion level, then an interview should be brought gently to a close, since nothing will be learned from subjecting her or him to a longer ordeal. Where, on the other hand, the purpose of the test is to see what level the candidate is at, in an interview the tester has to begin by guessing what this level is on the basis of early responses. The interview is then conducted at that level, either providing confirmatory evidence

or revealing that the initial guess is inaccurate. In the latter case the level is shifted up or down until it becomes clear what the candidate's level is. A second tester, whose main role is to assess the candidate's performance, can elicit responses at a different level if it is suspected that the principal interviewer may be mistaken.

9. Do not talk too much. There is an unfortunate tendency for interviewers to talk too much, not giving enough talking time to candidates. Avoid the temptation to make lengthy or repeated explanations of something that the candidate has misunderstood.
10. Select interviewers carefully and train them. Successful interviewing is by no means easy and not everyone has great aptitude for it. Interviewers need to be sympathetic and flexible characters, with a good command of the language themselves. But even the most apt need training. What follows is the outline of a possible four-stage training programme for interviewers, where interviewing is carried out as recommended above, with two interviewers.

Stage 1 Background and overview

- Trainees are given background on the interview.
- Trainees are given a copy of the handbook and taken through its contents.
- The structure of the interview is described.
- A video of a typical interview is shown.
- Trainees are asked to study the handbook before the second stage of the training.

Stage 2 Assigning candidates to levels

- Queries arising from reading the handbook are answered.
- A set of calibrated videos is shown.
- After each video, trainees are asked to write down the levels to which they assign the candidate according to the level descriptions and the analytic scale, and to complete a questionnaire on the task. A discussion follows.
- All papers completed by trainees during this stage are kept as a record of their performance.

Stage 3 Conducting interviews

- Pairs of trainees conduct interviews, which are videoed.
- The other trainees watch the interview on a monitor in another room.
- After each interview, all trainees assign the candidate to a level and complete a questionnaire. These are then discussed.
- Each trainee will complete six interviews.

Stage 4 Assessment

- Procedures will be as in Stage 3, except that the performance of trainees will not be watched by other trainees. Nor will there be any discussion after each interview.
- Ensure valid and reliable scoring.
- Create appropriate scales for scoring.

As was said for tests of writing in the previous chapter, rating scales may be holistic or analytic. The advantages and disadvantages of the two approaches have already been discussed in the previous chapter. We begin by looking at the assessment criteria for *Cambridge English B2 First*. These will have been applied to candidates performing the tasks presented above.

B2	Grammar and Vocabulary	Discourse Management	Pronunciation	Interactive Communication
5	Shows a good degree of control of a range of simple and some complex grammatical forms. Uses a range of appropriate vocabulary to give and exchange views on a wide range of familiar topics.	Produces extended stretches of language with very little hesitation. Contributions are relevant and there is a clear organisation of ideas. Uses a range of cohesive devices and discourse markers.	Is intelligible. Intonation is appropriate. Sentence and word stress is accurately placed. Individual sounds are articulated clearly.	Initiates and responds appropriately, linking contributions to those of other speakers. Maintains and develops the interaction and negotiates towards an outcome.
4	<i>Performance shares features of Bands 3 and 5.</i>			
3	Shows a good degree of control of simple grammatical forms, and attempts some complex grammatical forms. Uses a range of appropriate vocabulary to give and exchange views on a range of familiar topics.	Produces extended stretches of language despite some hesitation. Contributions are relevant and there is very little repetition. Uses a range of cohesive devices.	Is intelligible. Intonation is generally appropriate. Sentence and word stress is generally accurately placed. Individual sounds are generally articulated clearly.	Initiates and responds appropriately. Maintains and develops the interaction and negotiates towards an outcome with very little support.
2	<i>Performance shares features of Bands 1 and 3.</i>			
1	Shows a good degree of control of simple grammatical forms. Uses a range of appropriate vocabulary when talking about everyday situations.	Produces responses which are extended beyond short phrases, despite hesitation. Contributions are mostly relevant, despite some repetition. Uses basic cohesive devices.	Is mostly intelligible, and has some control of phonological features at both utterance and word levels.	Initiates and responds appropriately. Keeps the interaction going with very little prompting and support.
0	<i>Performance below Band 1.</i>			

Cambridge English B2 First differs from the ILR descriptors below in that it does specify functions separately.

The Interagency Language Roundtable (ILR) speaking levels go from 0 (zero) to 5 (expert speaker), with a plus indicating a level intermediate between two 'whole number' levels. Levels 2, 2+ and 3 follow.

SPEAKING 2 (LIMITED WORKING PROFICIENCY)

Able to satisfy routine social demands and limited work requirements. Can handle routine work-related interactions that are limited in scope. In more complex and sophisticated work-related tasks, language usage generally disturbs the native speaker. Can handle with confidence, but not with facility, most normal, high-frequency social conversational situations including extensive, but casual conversations about current events, as well as work, family, and autobiographical information. The individual can get the gist of most

everyday conversations but has some difficulty understanding native speakers in situations that require specialized or sophisticated knowledge. The individual's utterances are minimally cohesive. Linguistic structure is usually not very elaborate and not thoroughly controlled; errors are frequent. Vocabulary use is appropriate for high-frequency utterances, but unusual or imprecise elsewhere.

Examples: While these interactions will vary widely from individual to individual, the individual can typically ask and answer predictable questions in the workplace and give straightforward instructions to subordinates. Additionally, the individual can participate in personal and accommodation-type interactions with elaboration and facility; that is, can give and understand complicated, detailed, and extensive directions and make non-routine changes in travel and accommodation arrangements. Simple structures and basic grammatical relations are typically controlled; however, there are areas of weakness. In the commonly taught languages, these may be simple markings such as plurals, articles, linking words, and negatives or more complex structures such as tense/aspect usage, case morphology, passive constructions, word order, and embedding.

SPEAKING 2+ (LIMITED WORKING PROFICIENCY, PLUS)

Able to satisfy most work requirements with language usage that is often, but not always, acceptable and effective. The individual shows considerable ability to communicate effectively on topics relating to particular interests and special fields of competence.

Often shows a high degree of fluency and ease of speech, yet when under tension or pressure, the ability to use the language effectively may deteriorate. Comprehension of normal native speech is typically nearly complete. The individual may miss cultural and local references and may require a native speaker to adjust to his/her limitations in some ways. Native speakers often perceive the individual's speech to contain awkward or inaccurate phrasing of ideas, mistaken time, space, and person references, or to be in some way inappropriate, if not strictly incorrect.

Examples: Typically the individual can participate in most social, formal, and informal interactions; but limitations either in range of contexts, types of tasks, or level of accuracy hinder effectiveness. The individual may be ill at ease with the use of the language either in social interaction or in speaking at length in professional contexts. He/she is generally strong in either structural precision or vocabulary, but not in both. Weakness or unevenness in one of the foregoing, or in pronunciation, occasionally results in miscommunication. Normally controls, but cannot always easily produce, general vocabulary. Discourse is often incohesive.

SPEAKING 3 (GENERAL PROFESSIONAL PROFICIENCY)

Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics. Nevertheless, the individual's limitations generally restrict the professional contexts of language use to matters of shared knowledge and/or international convention. Discourse is cohesive. The individual uses the language acceptably, but with some noticeable imperfections; yet, errors virtually never interfere with understanding and rarely disturb the native speaker. The individual can effectively combine structure and vocabulary to convey his/her meaning accurately. The individual

speaks readily and fills pauses suitably. In face-to-face conversation with natives speaking the standard dialect at a normal rate of speech, comprehension is quite complete. Although cultural references, proverbs, and the implications of nuances and idiom may not be fully understood, the individual can easily repair the conversation. Pronunciation may be obviously foreign. Individual sounds are accurate; but stress, intonation, and pitch control may be faulty.

Examples: Can typically discuss particular interests and special fields of competence with reasonable ease. Can use the language as part of normal professional duties such as answering objections, clarifying points, justifying decisions, understanding the essence of challenges, stating and defending policy, conducting meetings, delivering briefings, or other extended and elaborate informative monologues. Can reliably elicit information and informed opinion from native speakers. Structural inaccuracy is rarely the major cause of misunderstanding. Use of structural devices is flexible and elaborate. Without searching for words or phrases, individual uses the language clearly and relatively naturally to elaborate concepts freely and make ideas easily understandable to native speakers. Errors occur in low-frequency and highly complex structures.

It was said that holistic and analytic scales can be used as a check on each other. An example of this in oral testing is the American FSI (Foreign Service Institute) interview procedure³ which requires the two testers concerned in each interview both to assign candidates to a level holistically and to rate them on a six-point scale for each of the following: accent, grammar, vocabulary, fluency, comprehension. These ratings are then weighted and totalled. The resultant score is then looked up in a table which converts scores into the holistically described levels. The converted score should give the same level as the one to which the candidate was first assigned. If not, the testers will have to reconsider whether their first assignments were correct. The weightings and the conversion tables are based on research which revealed a very high level of agreement between holistic and analytic scoring. Having used this system when testing bank staff, we can attest to its efficacy. For the reader's interest we reproduce the rating scales and the weighting table. It must be remembered, however, that these were developed for a particular purpose and should not be expected to work well in a significantly different situation without modification. It is perhaps also worth reminding the reader that the use of a 'native-speaker' standard against which to judge performance is generally regarded as inappropriate, as we noted in Chapter 7. The reader who wishes to use the procedure should feel free to make changes to the terminology.

³ We understand that the FSI no longer tests speaking ability in the way that it did. However, we have found the methods described in their 'Testing Kit', which also includes both holistic and analytic scales, very useful when testing the language ability of professional people in various situations.

PROFICIENCY DESCRIPTIONS

Accent

1. Pronunciation frequently unintelligible.
2. Frequent gross errors and a very heavy accent make understanding difficult, require frequent repetition.
3. "Foreign accent" requires concentrated listening, and mispronunciations lead to occasional misunderstanding and apparent errors in grammar or vocabulary.
4. Marked "foreign accent" and occasional mispronunciations which do not interfere with understanding.
5. No conspicuous mispronunciations, but would not be taken for a native speaker.
6. Native pronunciation, with no trace of "foreign accent."

Grammar

1. Grammar almost entirely inaccurate except in stock phrases.
2. Constant errors showing control of very few major patterns and frequently preventing communication.
3. Frequent errors showing some major patterns uncontrolled and causing occasional irritation and misunderstanding.
4. Occasional errors showing imperfect control of some patterns but no weakness that causes misunderstanding.
5. Few errors, with no patterns of failure.
6. No more than two errors during the interview.

Vocabulary

1. Vocabulary inadequate for even the simplest conversation.
2. Vocabulary limited to basic personal and survival areas (time, food, transportation, family, etc.).
3. Choice of words sometimes inaccurate, limitations of vocabulary prevent discussion of some common professional and social topics.
4. Professional vocabulary adequate to discuss special interests; general vocabulary permits discussion of any non-technical subject with some circumlocutions.
5. Professional vocabulary broad and precise; general vocabulary adequate to cope with complex practical problems and varied social situations.
6. Vocabulary apparently as accurate and extensive as that of an educated native speaker.

Fluency

1. Speech is so halting and fragmentary that conversation is virtually impossible.
2. Speech is very slow and uneven except for short or routine sentences.
3. Speech is frequently hesitant and jerky; sentences may be left uncompleted.
4. Speech is occasionally hesitant, with some unevenness caused by rephrasing and groping for words.
5. Speech is effortless and smooth, but perceptively non-native in speed and evenness.
6. Speech on all professional and general topics as effortless and smooth as a native speaker's.

Comprehension

1. Understands too little for the simplest type of conversation.
2. Understands only slow, very simple speech on common social and touristic topics; requires constant repetition and rephrasing.
3. Understands careful, somewhat simplified speech when engaged in a dialogue, but may require considerable repetition and re-phrasing.
4. Understands quite well normal educated speech when engaged in a dialogue, but requires occasional repetition or rephrasing.
5. Understands everything in normal educated conversation except for very colloquial or low-frequency items, or exceptionally rapid or slurred speech.
6. Understands everything in both formal and colloquial speech to be expected of an educated native speaker.

WEIGHTING TABLE

	1	2	3	4	5	6	(A)
Accent	0	1	2	2	3	4	_____
Grammar	6	12	18	24	30	36	_____
Vocabulary	4	8	12	16	20	24	_____
Fluency	2	4	6	8	10	12	_____
Comprehension	4	8	12	15	19	23	_____
						Total	_____

Note the relative weightings for the various components.
The total of the weighted scores is then looked up in the following table, which converts it into a rating on a scale 0-4+.

CONVERSION TABLE

Score	Rating	Score	Rating	Score	Rating
16-25	0+	43-52	2	73-82	3+
26-32	1	53-62	2+	83-92	4
33-42	1+	63-72	3	93-99	4+

(Adams and Frith 1979)

Where analytic scales of this kind are used to the exclusion of holistic scales, the question arises (as with the testing of writing) as to what pattern of scores (for an individual candidate) should be regarded as satisfactory. This is really the same problem (though in a more obvious form) as the failure of individuals to fit holistic descriptions. Once again it is a matter of agreeing, on the basis of experience, what failures to reach the expected standard on particular parameters are acceptable.

The advice on creating rating scales given in the previous chapter is equally relevant here:

Calibrate the scale to be used

Generally the same procedures are followed in calibrating speaking scales as were described for writing scales, with the obvious difference that video-recordings are used rather than pieces of written work.

Train scorers (as opposed to interviewers)

The training of interviewers has already been outlined. Where raters are used to score interviews without acting as interviewers themselves, or are involved in the rating of responses to audio- or video-recorded stimuli, the same methods can be used as for the training of raters of written work.

Follow acceptable scoring procedures

Again, the advice that one would want to offer here is very much the same as has already been given in the previous chapter. Perhaps the only addition to be made is that great care must be taken to ignore personal qualities of the candidates that are irrelevant to an assessment of their language ability. We remember well the occasion when raters quite seriously underestimated the ability of one young woman who had dyed her hair blonde. In a speaking test it can be difficult to separate such features as pleasantness, confidence, or even someone's choice of outfit, from their language ability – but one must try!

Conclusion

The accurate measurement of speaking ability is not easy. It takes considerable time and effort, including training, to obtain valid and reliable results. Nevertheless, where a test is high-stakes, or backwash is an important consideration, the investment of such time and effort may be considered necessary. Readers are reminded that the appropriateness of content, of rating scales levels, and of elicitation techniques used in oral testing will depend upon the needs of individual institutions or organisations.



READER ACTIVITIES

These two activities are best carried out with colleagues.

Activity A

1. Visit the Trinity College website and familiarise yourself with the performance descriptors for the *Graded Examination in Spoken English (GESE)*.
2. Now watch the sample *GESE* videos on the same website and assign a grade (A–D) to each candidate.
3. Look at the 'Marks and Rationale' document and compare the designated grades with those that you gave the candidates. What do you notice?

Activity B

1. For a group of students that you are familiar with, prepare a holistic rating scale (five bands) appropriate to their range of ability. From your knowledge of the students, place each of them on this scale.

2. Choose three methods of elicitation (for example, role play, group discussion, interview). Design a test in which each of these methods is used for five to ten minutes.
3. Administer the test to a sample of the students you first had in mind.
4. Note problems in administration and scoring. How would you avoid them?
5. For each student who takes the test, compare scores on the different tasks. Do different scores represent real differences of ability between tasks? How do the scores compare with your original ratings of the students?



FURTHER READING

General

Fulcher (2003) looks at the testing of second language speaking from both theoretical and practical perspectives. O'Sullivan (2012a) is a very approachable summary of the issues related to task types and performance scoring.

Pair and group interaction

Van Moere (2006) investigates the validity of a group oral test and whether such tests are suitable for high-stakes assessment. Nakatsuhara (2011) uses conversation analysis to explore the effect of individual test-takers' levels of extraversion on the conversation style of a group. Ockey (2009) investigates the relationship between test-takers' levels of assertiveness and their score in group oral assessment. *Language Testing* 26, 3 (2009) is a special issue on pairwork in L2 assessment.

Interviewers and raters

Brown (2003) investigates the influence that interviewer differences can have on the elicited performance of test-takers. Carey et al. (2011) investigate how raters' familiarity with the L1 of test-takers can influence their perceptions of the performance. O'Loughlin (2002) examines the issue of gender in oral interviews, in terms of both the rating process and the effect on the discourse pattern during the interview. Zhang and Elder (2011) investigate differences between native speaker and non-native speaker raters and how the two groups perceive oral proficiency. Lazaraton (1996) examines the kinds of linguistic and interactional support which interlocutors may give to candidates. Lumley and McNamara (1995) report on a study into rater bias in oral testing. Wigglesworth (1993) shows how bias in raters can be detected and how raters can improve when their bias is brought to their attention.

Semi-direct and automated testing

O'Loughlin (2001) explores the equivalence of direct and semi-direct tests of speaking. Bernstein et al. (2010) review the validity of automated speaking tests against that of traditional oral proficiency interviews. Pearson Assessments (online) report on the use of automated scoring of speaking.

Communicative competence

Roever and Kasper (2018) argue for interactional competence to be incorporated into speaking assessment. Roever (2011) reviews existing tests of pragmatic competence and makes suggestions for future pragmatics tests. Youn (2015) investigates the effectiveness of role play activities in assessing pragmatic competence.

11

Testing reading

This chapter begins by considering how we should specify what candidates can be expected to do, and then goes on to make suggestions for setting appropriate test tasks.

Specifying what the candidate should be able to do

Operations

The testing of reading ability seems deceptively straightforward when it is compared to, say, the testing of speaking ability. You take a passage, ask some questions about it, and there you are. But while it is true that you can very quickly construct a reading test, it may not be a very good test, and it may not measure what you want it to measure.

The basic problem is that the exercise of receptive skills does not necessarily, or usually, manifest itself directly in overt behaviour. When people write and speak, we see and hear; when they read and listen, there will often be nothing to observe. The challenge for the language tester is to set tasks which will not only cause the candidate to exercise reading (or listening) skills, but will also result in behaviour that will demonstrate the successful use of those skills. There are two parts to this problem. First, there is uncertainty about the skills which may be involved in reading and which, for various reasons, language testers are interested in measuring; many have been hypothesised but few have been unequivocally demonstrated to exist. Second, even if we believe in the existence of a particular skill, it is still difficult to know whether an item has succeeded in measuring it.

The proper response to this problem is not to resort to the simplistic approach to the testing of reading outlined in the first paragraph, while we wait for confirmation that the skills we think exist actually do. We believe these skills exist because we are readers ourselves and are aware of at least some of them. We know that, depending on our purpose in reading and the kind of text we are dealing with, we may read in quite different ways. On one occasion we may read slowly and carefully, word by word, to follow, say, a philosophical argument. Another time we may flit from page to page, pausing only a few seconds on each, to get the gist of something. At yet another time we may look quickly down a column of text, searching for a particular piece of information. There is little doubt that accomplished readers are skilled in adapting the way they read according to purpose and

text. This being so, we see no difficulty in including these different kinds of reading in the specifications of a test.

If we reflect on our reading, we become conscious of other skills we have. Few of us will know the meaning of every word we ever meet, yet we can often infer the meaning of a word from its context. Similarly, as we read, we are continually making inferences about people, things and events. If, for example, we read that someone has spent an evening in a pub and that he then staggers home, we may infer that he staggers because of what he has drunk. (We realise that he could have been an innocent footballer who had been kicked on the ankle in a match and then gone to the pub to drink lemonade, but we didn't say that all our inferences were correct.)

It would not be helpful to continue giving examples of the reading skills we know we have. The point is that we do know they exist. The fact that not all of them have had their existence confirmed by research is not a reason to exclude them from our specifications, and thereby from our tests. The question is: Will it be useful to include them in our test? The answer might be thought to depend at least to some extent on the purpose of the test. If it is a diagnostic test which attempts to identify in detail the strengths and weaknesses in learners' reading abilities, the answer is certainly yes. If it is an achievement test, where the development of these skills is an objective of the course, the answer must again be yes. If it is a placement test, where a rough-and-ready indication of reading ability is enough, or a proficiency test where an 'overall' measure of reading ability is sufficient, one might expect the answer to be no. But the answer 'no' invites a further question. If we are not going to test these skills, what are we going to test? Each of the questions that were referred to in the first paragraph must be testing *something*. If our items are going to test something, surely on grounds of validity, in a test of overall ability, we should try to test a sample of all the skills that are involved in reading and are relevant to our purpose. This is what we would recommend.

Of course, the weasel words in the previous sentence are 'relevant to our purpose'. For beginners, there may be an argument for including in a diagnostic test items which test the ability to distinguish between letters (e.g. between *b* and *d*). But normally this ability will be tested indirectly through higher-level items. The same is true for grammar and vocabulary. They are both tested indirectly in every reading test, but the place for grammar and vocabulary items is, we would say, in grammar and vocabulary tests. For that reason we will not discuss them further in this chapter.

To be consistent with our general framework for specifications, we will refer to the skills that readers perform when reading a text as *operations*. In the boxes that follow are checklists (not meant to be exhaustive) which it is thought the reader of this book may find useful. Note the distinction, based on differences of purpose, between expeditious (quick and efficient) reading and slow and careful reading. There has been a tendency in the

past for expeditious reading to be given less prominence in tests than it deserves. The backwash effect of this is that many students have not been trained to read quickly and efficiently. This is a considerable disadvantage when, for example, they study overseas and are expected to read extensively in very limited periods of time. Another example of harmful backwash!



EXPEDITIOUS READING OPERATIONS

Surveying

The candidate can decide the relevance of a text (or part of a text) to their needs by looking at the author, sub-headings, graphics, etc.

Skimming

The candidate can:

- obtain main ideas and discourse topics quickly and efficiently;
- establish quickly the structure of a text.

Search reading

The candidate can quickly find information on a predetermined topic.

Scanning

The candidate can quickly find:

- specific words or phrases;
- figures, percentages;
- specific items in an index;
- specific names in a bibliography or a set of references.

Note that any serious testing of expeditious reading will require candidates to respond to items without having time to read the full contents of a passage.



CAREFUL READING OPERATIONS

- identify pronominal reference;
- identify discourse markers;
- interpret complex sentences;
- interpret topic sentences;
- outline logical organisation of a text;
- outline the development of an argument;
- distinguish general statements from examples;
- identify explicitly stated main ideas;
- identify implicitly stated main ideas;
- recognise writer's intention;

- recognise the attitudes and emotions of the writer;
- identify addressee or audience for a text;
- identify what kind of text is involved (e.g. editorial, diary, etc.);
- distinguish fact from opinion;
- distinguish hypothesis from fact;
- distinguish fact from rumour or hearsay.

Make inferences:

- infer the meaning of an unknown word from context;
- make propositional informational inferences, answering questions beginning with *who*, *when*, *what*;
- make propositional explanatory inferences concerned with motivation, cause, consequence and enablement, answering questions beginning with *why*, *how*;
- make pragmatic inferences.

The different kinds of inference described above deserve comment. Propositional inferences are those which do not depend on information from outside the text. For example, if John is Mary's brother, we can infer that Mary is John's sister (if it is also clear from the text that Mary is female). Another example: if we read the following, we can infer that Harry was working at her studies, not at the fish and chip shop. *Harry worked as hard as she had ever done in her life. When the exam results came out, nobody was surprised that she came top of the class.*

Pragmatic inferences are those where we have to combine information from the text with knowledge from outside the text. We may read, for example: *It took them twenty minutes by road to get from Reading to Heathrow Airport.* In order to infer that they travelled very quickly, we have to know that Reading and Heathrow Airport are not close by each other. The fact that many readers will not know this allows us to make the point that where the ability to make pragmatic inferences is to be tested, the knowledge that is needed from outside the text must be knowledge which all the candidates can be assumed to have¹.

Texts

Texts that candidates are expected to be able to deal with can be specified along a number of parameters: type, form, graphic features, topic, style, intended readership, length, readability or difficulty, range of vocabulary and grammatical structure.

¹ It has to be admitted that the distinction between propositional and pragmatic inferences is not watertight. In a sense all inferences are pragmatic: even being able to infer, say, that a man born in 1941 will have his ninetieth birthday in 2031 (if he lives that long) depends on knowledge of arithmetic, it could be argued. However, the distinction remains useful when we are constructing reading test items. Competent readers integrate information from the text into their knowledge of the world.

Text types include: textbooks, handouts, articles (in newspapers, journals or magazines), poems/verse, encyclopaedia entries, text messages, tweets, dictionary entries, web pages, blogs, leaflets, letters, forms, diary entries, maps or plans, advertisements, postcards, social media posts, timetables, novels (extracts) and short stories, reviews, manuals, computer Help systems, notices and signs.

Text forms include: description, exposition, argumentation, instruction, narration. (These can be broken down further if it is thought appropriate: e.g. expository texts could include outlines, summaries, etc.)

Graphic features include: tables, charts, diagrams, cartoons, illustrations, infographics.

Topics may be listed or defined in a general way (such as non-technical, non-specialist) or in relation to a set of candidates whose background is known (such as those familiar to the students).

Style may be specified in terms of formality.

Intended readership can be quite specific (e.g. expert speaking science undergraduate students) or more general (e.g. young expert speakers).

Length is usually expressed in number of words. The specified length will normally vary according to the level of the candidates and whether one is testing expeditious or careful reading (although a single long text could be used for both).

Readability is an objective, but not necessarily very valid, measure of the difficulty of a text. Where this is not used, expert judgements may be relied on.

Range of vocabulary may be indicated by a complete list of words (as for the Cambridge tests for young learners), by reference either to a word list or to indications of frequency in a learners' dictionary. The free online resource, *English Vocabulary Profile (EVP)* is particularly useful here. Range may be expressed more generally (e.g. non-technical, except where explained in the text).

Range of grammar may be a list of structures, or a reference to those to be found in a course book or (possibly parts of) a grammar of the language.

The reason for specifying texts in such detail is that we want the texts included in a test to be representative of the texts candidates should be able to read successfully. This is partly a matter of content validity but also relates to backwash. The appearance in the test of only a limited range of texts will encourage the reading of a narrow range by potential candidates.

It is worth mentioning authenticity at this point. Whether or not authentic texts (intended for expert speakers) are to be used will depend at least in part on what the items based on them are intended to measure.

Speed

Reading speed may be expressed in words per minute. Different speeds will be expected for careful and expeditious reading. In the case of the latter, the candidate is, of course, not expected to read all of the words. The expected speed of reading will combine with the number and difficulty of items to determine the amount of time needed for the test, or part of it. While research has suggested that 250 words per minute is a reasonable *target* reading speed for fluent second language reading, expectations for particular groups of learners will vary according to their general level of proficiency, the nature of the text and the tasks which they are asked to perform. Observation of learners reading texts is the best guide to setting a reading speed.

Criterion level of performance

In norm-referenced testing our interest is in seeing how candidates perform by comparison with each other. There is no need to specify criterion levels of performance before tests are constructed, or even before they are administered. This book, however, encourages a broad criterion-referenced approach to language testing. In the case of the testing of writing, as we saw in Chapter 9, it is possible to describe levels of writing ability that candidates have to attain. While this would not satisfy everyone's definition of criterion-referencing, it is very much in the spirit of that form of testing, and would promise to bring the benefits claimed for criterion-referenced testing.

Setting criterion levels for receptive skills is more problematical. Traditional pass marks expressed in percentages (40 percent? 50 percent? 60 percent?) are hardly helpful, since there seems no way of providing a direct interpretation of such a score. To our minds, the best way to proceed is to use the test tasks themselves to define the level. All of the items (and so the tasks that they require the candidate to perform) should be within the capabilities of anyone to whom we are prepared to give a pass. In other words, in order to pass, a candidate should be expected, in principle, to score 100 percent. But since we know that human performance is not so reliable, we can set the actual cutting point rather lower, say at the 80 percent level. In order to distinguish between candidates of different levels of ability, more than one test may be required.

As part of the development (and validation) of a reading test, one might wish to compare performance on the test with the rating of candidates' reading ability using scales like those of ACTFL or the ILR. This would be most appropriate where performance in the productive skills is being assessed according to those scales and some equivalence between tests of the different skills is being sought. Similarly, performance on the test may be compared with candidates' ability assessed in terms of *CEFR/ALTE* 'Can do' statements.

Setting the tasks

Selecting texts

Successful choice of texts depends ultimately on experience, judgement and a certain amount of common sense. Clearly these are not qualities that a handbook can provide; practice is necessary. It is nevertheless possible to offer useful advice. While these points may seem rather obvious, they are often overlooked.

1. Keep specifications constantly in mind and try to select as representative a sample as possible. Do not repeatedly select texts of a particular kind simply because they are readily available.
2. Choose texts of appropriate lengths. Expeditious reading tests may call for passages of up to 2,000 words or more. Detailed reading can be tested using passages of just a few sentences.
3. In order to obtain both content validity and acceptable reliability, include as many passages as possible in a test, thereby giving candidates a good number of fresh starts. Considerations of practicality will inevitably impose constraints on this, especially where scanning or skimming is to be tested.
4. In order to test search reading, look for passages which contain plenty of discrete pieces of information.
5. For scanning, find texts which have the specified elements that have to be scanned for.
6. To test the ability to quickly establish the structure of a text, make sure that the text has a clearly recognisable structure. (It's surprising how many texts lack this quality.)
7. Choose texts that will interest candidates but which will not over-excite or disturb them. A text about cancer, for example, is almost certainly going to be distressing to some candidates.
8. Avoid texts made up of information that may be part of candidates' general knowledge. It may be difficult not to write items to which correct responses are available to some candidates without reading the passage. On a reading test we encountered once, one of us was able to answer eight out of 11 items without reading the text on which they were based. The topic of the text was rust in cars, an area in which we had had extensive experience.
9. Assuming that it is only reading ability that is being tested, do not choose texts that are too culturally laden.

10. Do not use texts that students have already read (or even close approximations to them). This happens surprisingly often.

Writing items

The aim must be to write items that will measure the ability in which we are interested, that will elicit reliable behaviour from candidates, and that will permit highly reliable scoring. Since the act of reading does not in itself demonstrate its successful performance, we need to set tasks that will involve candidates in providing evidence of successful reading.

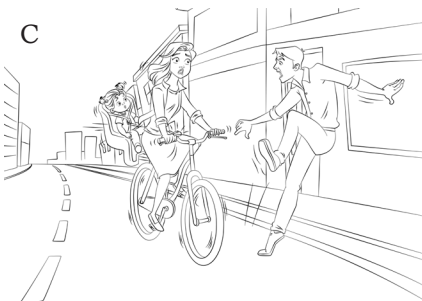
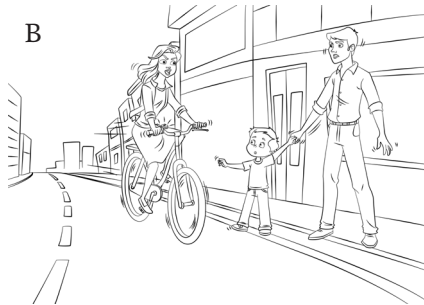
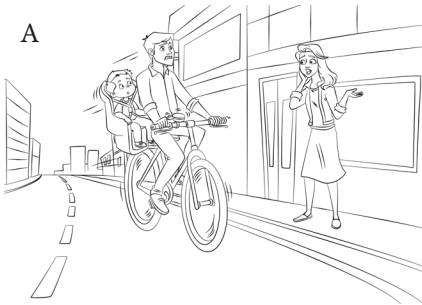
Possible techniques

It is important that the techniques used should interfere as little as possible with the reading itself, and that they should not add a significantly difficult task on top of reading. This is one reason for being wary of requiring candidates to write answers, particularly in the language of the text. They may read perfectly well but difficulties in writing may prevent them demonstrating this. Possible solutions to this problem include:

Multiple choice

The candidate provides evidence of successful reading by making a mark against one out of a number of alternatives. The superficial attraction of this technique is outweighed in institutional testing by the various problems enumerated in Chapter 8. This is true whether the alternative responses are written or take the form of illustrations, as in the following:

Choose the picture (A, B, C or D) that the following sentence describes:
The man with the child was shouted at by the woman on the bike.



It has already been pointed out that *True/False* items, which are to be found in many tests, are simply a variety of multiple choice, with only one distractor and a 50 percent probability of choosing the correct response by chance! Having a 'not applicable' or 'we don't know' category adds a second 'distractor' and reduces the likelihood of guessing correctly to 33 percent.

Short answer

The best short answer questions are those with a unique correct response, for example:

In which city do the people described in the 'Urban Villagers' live?
to which there is only one possible correct response, e.g. *Bombay*.

The response may be a single word or something slightly longer (e.g. *China and Japan; American women*).

The short answer technique works well for testing the ability to identify referents. An example (based on the newspaper article about the re-creation of ancient foods on page 152) is:

What does the word 'she' (line 53) refer to?

Care has to be taken that the precise referent is to be found in the text. It may be necessary on occasion to change the text slightly for this condition to be met.

The technique also works well for testing the ability to predict the meaning of unknown words from context. An example (also based on the ancient foods article) is:

Find a single word in the passage (between lines 10 and 20) which has the same meaning as 'minute opening or passage'. (The word in the passage may have an ending like *-s*, *-tion*, *-ing*, *-ed*, etc.)

The short answer technique can be used to test the ability to make various distinctions, such as that between fact and opinion. For example:

Basing your answers on the text, mark each of the following sentences as FACT or OPINION by writing F or O in the correct space on your answer sheet. You must get all three correct to obtain credit.

1. Farm owners are deliberately neglecting their land.
2. The majority of young men who move to the cities are successful.
3. There are already enough farms under government control.

Because of the requirement that all three responses are correct, guessing has a limited effect in such items.

Scanning can be tested with the short answer technique:

Which town listed in Table 4 has the largest population?

According to the index, on which page will you learn about Nabokov's interest in butterflies?

The short answer technique can also be used to write items related to the structure of a text. For example:

There are five sections in the paper. In which section do the writers deal with:

- choice of language in relation to national unity [Section]
- the effects of a colonial language on local culture [Section]
- the choice of a colonial language by people in their fight for liberation [Section]
- practical difficulties in using local languages for education [Section]
- the relationship between power and language [Section]

Again, guessing is possible here, but the probabilities are lower than with straightforward multiple choice.

A similar example is shown below from *Cambridge Complete First 2nd edition Student's Book*²:

1 You are going to read an extract from a magazine article. Six paragraphs have been removed from the extract. Choose from the paragraphs A–G the one which fits each gap 1–6. There is one extra paragraph which you do not need to use.

2 Work in pairs. Discuss the words/phrases which helped you to decide what fits where.

Is your glass half full or half empty?

Are you happy? Did you open the curtains this morning, see that it's yet another day of sunshine and bounce out of bed? Or are you the kind of person who sees the sun and starts worrying about getting sunburnt and the problems it may cause for gardeners?

1

But a television documentary, which is to be broadcast next week, suggests that in fact they play only a very small part and that you can, in fact, train yourself to have a more sunny attitude to life. It argues that it may indeed be simple to change negative people into positive ones.

2

Next week's programme is timely, because the happiness of individuals is something that policymakers have started to take very seriously indeed. Indeed, yesterday, a new charity called MindFull suggested that mental health should be taught in schools. And later this month, the Office for National Statistics (ONS) will publish its National Well-being report. This will draw on a number of studies which suggest that our positivity has an impact on our health and our educational achievements.

3

In other words, being happy could add years to your life. It doesn't just benefit your health, either. Educational attainment, too, seems to be linked to attitude. Nick Baylis, a consultant psychologist, works with the pupils at a school in London that, five years ago, had very poor academic results. Now, 87% of its pupils are leaving school with good qualifications. Baylis believes that teaching both the staff and pupils 'well-being' and coping strategies was key to this success.

4

'Through monkeys, humans and lots of animals, the amount of activity in the front cortex does seem to be a good marker for positivity and negativity.' Positive people have a more active left frontal cortex; the presenter was found to have a substantially more active right frontal cortex – proving his assertion that he is one of life's pessimists. 'When I look into the future, I see all the things that are going to go wrong, rather than the things that will probably go right,' he says. He also suffers from insomnia. Professor Fox is among a growing number of psychologists, however, who believe that he and others like him can change this brain asymmetry and thus their personality through a series of exercises.

5

It seems simple. But surely, trying to pick out a smiling expression isn't going to make me more optimistic. Professor Fox tells me: 'I was very sceptical when I got into this initially. But the task we used in the show has been used with kids with self-esteem issues. And it does seem to have very powerful effects. It's early days, but the signs are that it is definitely effective.'

6

Of course, many psychologists argue that relentless happiness is neither normal nor healthy. Professor Fox says: 'There are situations when things go wrong, and having a healthy dose of pessimism can be good. But the evidence shows that, broadly, having a positive attitude really does boost your well-being.'



² Note that this example is taken from an exam *preparation* book, hence the instruction to work in pairs, which of course would not be appropriate in a test proper.

A The most striking example comes from Oxford, Ohio, which in the 1970s conducted a study of its inhabitants, then aged over 50. So who has survived in good health? Those who had a positive outlook on their life and impending old age have lived, on average, 7.6 years longer than those with negative views.

B It worked for the presenter, who over a couple of months of exercising was able to recalibrate his brain. He says that he is sleeping better 'though I wouldn't call myself a heavy sleeper yet', and that he is more optimistic. So should we all be doing the exercises? 'I think anyone could do them, but I suspect a fair number who start then let it slide,' he says.

C If the show touches a nerve in the same way as last autumn's documentary by the same director about fasting – which kick-started the phenomenally popular 5:2 diet – many of us could soon be undertaking mental workouts in our lunch hour.

D Professor Fox gives her views on the subject in next week's programme, pointing out that the research has very significant implications for schools and for health professionals. 'However, more work needs to be done before the results can be considered conclusive.'

E The most basic one is called Cognitive Bias Modification. To do it, you look at a screen for 10 minutes every day over several weeks. During those minutes, a series of 15 faces are flashed up. All (except one) are either angry, upset or unhappy. You have to spot, and click on, the one happy face.

F For years, many scientists believed that your personality was predetermined. They were of the opinion that it was your genes which were responsible for whether you were an optimist or a pessimist.

G Next week's documentary will try to provide a physiological explanation for their achievements. For the programme, the presenter had his brain scanned by Professor Elaine Fox, a neuroscientist at Oxford and author of *Rainy Brain*, *Sunny Brain*. She says brain asymmetry is very closely linked to our personalities.

EXAM ADVICE

- Read the whole of the text first.
- Read through paragraphs A–G and notice the differences between them.
- Pay careful attention to connecting words throughout the text and paragraphs, as well as at the beginnings and ends of paragraphs.
- Consider each paragraph for every gap. Don't assume you have been correct in your previous answers as you go along!
- Read the whole of the text again when you have completed the task.
- Don't rely on matching up names, dates or numbers in the text and paragraphs just because they are the same or similar.
- Don't rely on matching up individual words or phrases in the text and the paragraphs just because they are the same or similar.

It should be noted that the scoring of 'sequencing' items of this kind can be problematical. If a candidate puts one element of the text out of sequence, it may cause others to be displaced and require complex decision-making on the part of the scorers.

One should be wary of writing short answer items where correct responses are not limited to a unique answer. Thus:

According to the author, what does the increase in divorce rates show about people's expectations of marriage and marriage partners?

might call for an answer like:

(They/Expectations) are greater (than in the past).

The danger is of course that a student who has the answer in his or her head after reading the relevant part of the passage may not be able to express it well (equally, the scorer may not be able to tell from the response that the student has arrived at the correct answer).

Gap filling

This technique is particularly useful in testing reading. It can be used any time that the required response is so complex that it may cause writing (and scoring) problems. If one wanted to know whether the candidate had

grasped the main idea(s) of the following paragraph, for instance, the item might be:

Complete the following, which is based on the paragraph below.

'Many universities in Europe used to insist that their students speak and write only _____. Now many of them accept _____ as an alternative, but not a _____ of the two.'

Until recently, many European universities and colleges not only taught EngEng but actually required it from their students; i.e. other varieties of standard English were not allowed. This was the result of a conscious decision, often, that some norm needed to be established and that confusion would arise if teachers offered conflicting models. Lately, however, many universities have come to relax this requirement, recognising that their students are as likely (if not more likely) to encounter NAmEng as EngEng, especially since some European students study for a time in North America. Many universities therefore now permit students to speak and write either EngEng or NAmEng, so long as they are consistent.

(Trudgill and Hannah 2017)

A possible weakness in this particular item is that the candidate has to provide one word (*mixture* or *combination*) which is not in the passage. In practice, however, it worked well.

Gap filling can be used to test the ability to recognise detail presented to support a main idea:

To support his claim that the Mafia is taking over Russia, the author points out that the sale of _____ in Moscow has increased by _____ percent over the last two years.

Gap filling can also be used for scanning items:

According to Figure 1, _____ percent of faculty members agree with the new rules.

Gap filling is also the basis for what has been called 'summary cloze'. In this technique, a reading passage is summarised by the tester, and then gaps are left in the summary for completion by the candidate. This is really an extension of the gap filling technique and shares its qualities. It permits the setting of several reliable but relevant items. Here is an extended reading example based on a newspaper article, with higher-level students in mind:

Below, you will find a newspaper article about the modern re-creation of ancient food, followed by a summary of the article.

The summary contains gaps. You must fill the gaps using **only words from the article**. There must be **ONLY ONE WORD** in each gap.

Ancient foods

During a 1954 BBC documentary about Tollund Man, the mysterious body of a hanged man discovered in a peat bog in Denmark, the noted archaeologist Sir Mortimer Wheeler ate a reconstruction of the 2,000-year-old's last meal. After tasting the porridge of barley, linseed and mustard seeds, he dabbed at his moustache and declared the mystery was solved: Tollund Man had killed himself rather than eat another spoonful.

Food reconstruction has come a long way since then. Last week Seamus Blackley, a scientist more famous for creating the Xbox, baked a sourdough loaf using yeast cultured from scrapings off 4,500-year-old Egyptian pottery at his home in California. The results, said one of his collaborators, Dr Serena Love, an Egyptologist from the University of Queensland, were "tangy and delicious". "I met Seamus for the first time today," she said. "As soon as I walked in the door he gave me a plate of bread." Blackley extracted samples from inside the ceramic pores of a clay pot from the Peabody Museum at Harvard University three weeks ago. Most are being examined by the third member of the team, Richard Bowman, a molecular biologist, but Blackley kept one to turn it into yeast to make bread. "Food puts you in touch with the humanity of the past," Love said. "That's a tactile thing, something that's visceral – you can actually experience the ancients, with at least one of the actual ingredients."

Ancient and historical foods are having a bit of a moment. The growing interest can be seen in the number of cookbooks available including *An Early Meal*, a *Viking Age Cookbook* by Daniel Serra and Hanna Tunberg and *Khazana* by Saliha Mahmood Ahmed with recipes inspired by the Mughal empire, as well as in the increasing number of food re-enactments. Graham Taylor's Potted History firm makes amphoras and Neolithic pottery for experimental archaeologists such as Sally Grainger who has investigated and made versions of garum, a Roman fish sauce, as well as Jill Hatch who cooks authentic Roman food for the Ermine Street Guard enthusiasts and similar groups. But those looking for original ingredients to recreate tastes of the past need to be cautious, says Professor Dorian Fuller, an archaeobotanist from University College London. "Yeast is everywhere. It's hard to know if something wasn't contaminated when it was dug out of the ground, or when it was put on a ship to Boston collecting yeasts along the way. These things haven't been kept in sterile conditions."

Because human diets have been founded on grains for millennia, beer, bread and porridge are the main focus of attempts to recreate truly ancient foods. "The latest study that came out in the '80s said grain made up about 70% of the daily diet of Romans, although I think

that's a little high," said Farrell Monaco, an archaeologist specialising in Roman culture who has worked in Pompeii and Herculaneum.

"Although I think that's a little high, bread and pulses were the two vehicles to get calories into the Roman daily diet." Pompeii has commercial bakeries on every street corner, she said. "And religion as well – bread was so valuable that you would offer it to the gods."

Monaco uses replicas of Roman and Greek kitchen tools to make dishes described by ancient writers such as Columella, Pliny and Cato: fig vinegar, moretum (salads), hypotrimma (a sweet paste) and defrutum (a grape syrup) as well as panis quadratus, a round loaf that has been excavated at many sites around Vesuvius. She believes making ancient food with original techniques is a vital archaeological tool. "To use your hand, your eyes, nose, tastebuds, to labour over something, to use a handmill to make a loaf of bread, so you understand how much labour and sweat went into making it – you start to understand how much value it had."

Summary

In a television documentary in 1954, an archaeologist made a joke, saying that a man had killed himself 2,000 years ago rather than eat any more of his _____, the remains of which had been found in his body.

Times have changed. Recently, scrapings were taken from 4,500 year old Egyptian _____. Most were kept for study by a molecular biologist, but one was retained to culture yeast, which was then used to bake a _____ loaf. An Egyptologist who tasted it said that it was tangy and delicious.

Growing interest in ancient foods is evidenced by the number of _____ which are being written, including two which provide recipes for Viking and Mughal empire inspired food. A firm called 'Potted History' makes amphoras and Neolithic pottery for archaeologists who want to make authentic ancient Roman food. At the same time, one archaeobotanist has warned that care should be exercised in such cookery, since yeast is everywhere and may _____ whatever is dug out of the ground.

The main focus of attempts to recreate ancient foods has been on beer, bread and porridge. This is because human diets have been based on _____ for thousands of years. A study in the 1980s claimed that about 70% of the _____ diet consisted of grain. Although she thinks that estimate to be a little high, Farrell Monaco, an archaeologist, admits that bread and pulses

were what provided Romans with their _____. Pompeii had bakeries on every street corner, she added. Monaco uses replica _____ to make dishes described by ancient writers. She believes that making bread in this way helps one understand the _____ it had for ancient peoples.

Information transfer

One way of minimising demands on candidates' writing ability is to require them to show successful completion of a reading task by supplying simple information in a table, following a route on a map, labelling a picture, and so on. As can be seen in the example below, from the *IELTS Academic* module, a single text may be used for more than one task (in this case, completing a table and labelling a picture).

[Note: This is an extract from an Academic Reading passage on the subject of dung beetles. The text preceding this extract gave some background facts about dung beetles, and went on to describe a decision to introduce non-native varieties to Australia.]

Introducing dung¹ beetles into a pasture is a simple process: approximately 1,500 beetles are released, a handful at a time, into fresh cow pats² in the cow pasture. The beetles immediately disappear beneath the pats digging and tunnelling and, if they successfully adapt to their new environment, soon become a permanent, self-sustaining part of the local ecology. In time they multiply and within three or four years the benefits to the pasture are obvious.

Dung beetles work from the inside of the pat so they are sheltered from predators such as birds and foxes. Most species burrow into the soil and bury dung in tunnels directly underneath the pats, which are hollowed out from within. Some large species originating from France excavate tunnels to a depth of approximately 30 cm below the dung pat. These beetles make sausage-shaped brood chambers along the tunnels. The shallowest tunnels belong to a much smaller Spanish species that buries dung in chambers that hang like fruit from the branches of a pear tree. South African beetles dig narrow tunnels of approximately 20 cm below the surface of the pat. Some surface-dwelling beetles, including a South African species, cut perfectly-shaped balls from the pat, which are rolled away and attached to the bases of plants.

For maximum dung burial in spring, summer and autumn, farmers require a variety of species with overlapping periods of activity. In the cooler environments of the state of Victoria, the large French species (2.5 cms long), is matched with smaller (half this size), temperate-climate Spanish species. The former are slow to recover from the winter cold and produce only one or two generations of offspring from late spring until autumn. The latter, which multiply rapidly in early spring, produce two to five generations annually. The South African ball-rolling species, being a sub-tropical beetle, prefers the climate of northern and coastal New South Wales where it commonly works with the South African tunneling species. In warmer climates, many species are active for longer periods of the year.

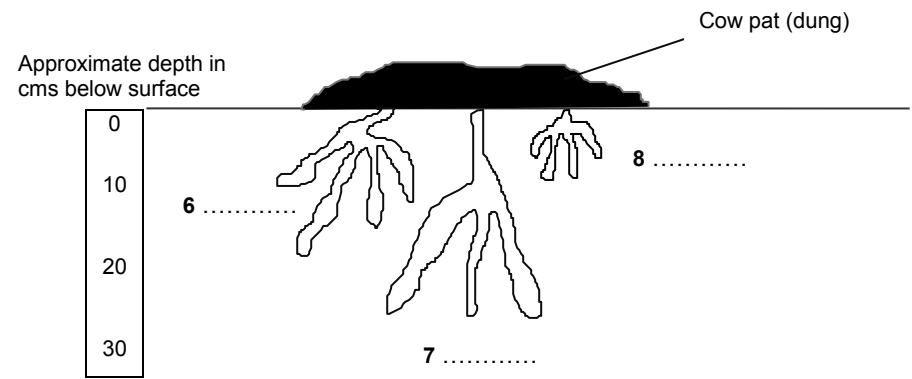
Glossary

1. dung: the droppings or excreta of animals
2. cow pats: droppings of cows

Questions 6 – 8

Label the tunnels on the diagram below using words from the box.

Write your answers in boxes 6-8 on your answer sheet.



Dung Beetle Types	
French	Spanish
Mediterranean	South African
Australian native	South African ball roller

Question 9 – 13

Complete the table below.

Choose **NO MORE THAN THREE WORDS** from the passage for each answer.

Write your answers in boxes 9-13 on your answer sheet.

Species	Size	Preferred climate	Complementary species	Start of active period	Number of generations per year
French	2.5 cm	cool	Spanish	late spring	1 - 2
Spanish	1.25 cm	9		10	11
South African ball roller		12	13		

Relatively few techniques have been presented in this section. This is because, in our view, few basic techniques are needed, and non-professional testers will benefit from concentrating on developing their skills within a limited range, always allowing for the possibility of modifying these techniques for particular purposes and in particular circumstances. Many professional testers appear to have got by with

just one – multiple choice! The more usual varieties of cloze and the C-Test technique (see Chapter 14) have been omitted because, while they obviously involve reading to quite a high degree, it is not clear that reading ability is all that they measure. This makes it all the harder to interpret scores on such tests in terms of criterial levels of performance.

Which language for items and responses?

The wording of reading test items is not meant to cause candidates any difficulties of comprehension. It should always be well within their capabilities, and less demanding than the text itself. In the same way, responses should make minimal demands on writing ability. Where candidates share a single native language, this can be used both for items and for responses. There is a danger, however, that items may provide some candidates with more information about the content of the text than they would have obtained from items in the foreign language.

Procedures for writing items

The starting point for writing items is a careful reading of the text, having the specified operations in mind. One should be asking oneself what a competent reader should derive from the text. Where relevant, a note should be taken of main points, interesting pieces of information, stages of argument, examples, and so on. The next step is to decide what tasks it is reasonable to expect candidates to be able to perform in relation to these. It is only then that draft items should be written. Paragraph numbers and line numbers should be added to the text if items need to make reference to these. The text and items should be presented to colleagues for moderation. Items and even the text may need modification. A moderation checklist follows:



MODERATION CHECKLIST

	YES	NO
1. Is the English of text and item grammatically correct?		
2. Is the English natural and acceptable?		
3. Is the item in accordance with specified parameters?		
4. Is the specified reading sub-skill necessary in order to respond correctly?		
5. (a) Multiple choice: Is there just one correct response?(b) Gap filling and summary cloze: Are there just one or two correct responses for each gap? (c) Short answer: Is the answer within productive abilities? Can it be scored validly and reliably? (d) Unique answer: Is there just one clear answer?		
6. Multiple choice: Are all the distractors likely to distract?		
7. Is the item economical?		
8. Is the key complete and correct?		

Practical advice on item writing

1. In a scanning test, present items in the order in which the answers can be found in the text. Not to do this introduces too much random variation and so lowers the test's reliability.
2. Do not write items for which the correct response can be found without understanding the text (unless that is an ability that you are testing!). Such items usually involve simply matching a string of words in the question with the same string in the text. Thus (around line 50 in the ancient foods passage, on page 153):

Who uses replicas of Roman and Greek kitchen tools to make dishes described by ancient writers such as Columella, Pliny and Cato?

Better might be:

Name the archaeologist who makes food described by Pliny and others.

Items that demand simple arithmetic can be useful here. We may learn in one sentence that before 2004 there had only been three hospital operations of a particular kind; in another sentence, that there have been 45 since. An item can ask how many such operations there have been to date, according to the article.

3. Do not include items that some candidates are likely to be able to answer from general knowledge without reading the text. For example:

Yeast is used in the making of _____

It is not necessary, however, to choose esoteric topics.

4. Make the items independent of each other; do not make a correct response on one item depend on another item being responded to correctly.

In the following example, the candidate who does not respond correctly to the first item is unlikely to be able to respond to the following two parts (the second of which uses the *Yes/No* technique). For such a candidate, b) and c) might as well not be there.

- a) Which man is suspected by the detective?
- b) What was the man wearing?
- c) Did the man attempt to escape?

However, complete independence is just about impossible in items that are related to the structure of a text.

5. Be prepared to make *minor* changes to the text to improve an item.

If you do this and are not an expert speaker, ask an expert speaker to look at the changed text.

A note on scoring

General advice on obtaining reliable scoring has already been given in Chapter 5. It is worth adding here, however, that in a reading test (or a listening test), errors of grammar, spelling or punctuation should not be penalised, provided that it is clear that the candidate has successfully performed the reading task which the item set. The function of a reading test is to test reading ability. To test productive skills at the same time (which is what happens when grammar, etc. are taken into account) simply makes the measurement of reading ability less valid.



READER ACTIVITIES

1. Following the procedures and advice given in the chapter, construct a six-item reading test based on the extract 'The secrets of happiness' on pages 159–160. (The passage comes from *Cambridge Complete First 2nd edition*.)
 - a. For each item, make a note of the skill(s) (including sub-skills) you believe it is testing. If possible, have colleagues take the test and provide critical comment. Try to improve the test. Again, if possible, administer the test to an appropriate group of students. Score the tests. Interview a few students as to how they arrived at correct responses. Did they use the particular sub-skills that you predicted they would?
 - b. Compare your questions with the ones in Appendix 3. Can you explain the differences in content and technique? Are there any items in the appendix that you might want to change? Why? How?
2. Do the sequencing item that is based on the text 'Is your glass half full or half empty?' In *Cambridge Complete First 2nd edition* on pages 149 and 150. Do you have any difficulties? If possible, get a number of students of appropriate ability to do the item, and then score their responses. Do you have any problems in scoring?
3. Write a set of short answer items with unique correct responses to replace the sequencing items that appear with the 'Is your glass half full or half empty?' text.
4. The following is an exercise designed to help students learn to cope with complex sentences. How successful would this form of exercise be as part of a reading test? What precisely would it test? Would you want to change the exercise in any way? If so, why and how? Could you make it non-multiple choice? If so, how?

The refusal of the government to consider alternatives to its policy on prisons, which was criticised by various human rights groups, both within the country and abroad, led to its downfall.

What is the subject of 'led to its downfall'?

- a. the refusal
- b. policy on prisons
- c. human rights groups
- d. the government

The secrets of happiness

Mihaly Csikszentmihalyi has devoted his life to studying happiness. He believes he has found the key.

- I've been fascinated by happiness most of my life. When I was a small boy, I noticed that though many of the adults around me were wealthy and educated, they were not always happy and this
 5 sometimes led them to behave in ways which I, as a child, thought strange. As a result of this, I decided to understand what happiness was and how best to achieve it. It was not surprising, then, that I decided to study psychology.
- 10 On arrival at the University of Chicago 50 years ago, I was disappointed to find that academic psychologists were trying to understand human behaviour by studying rats in a laboratory. I felt that there must be other more useful ways of
 15 learning how we think and feel. Although my original aim had been to achieve happiness for myself, I became more ambitious. I decided to build my career on trying to discover what made others happy also. I started out by studying
 20 creative people such as musicians, artists and athletes because they were people who devoted their lives to doing what they wanted to do, rather than things that just brought them financial rewards.
- 25 Later, I expanded the study by inventing a system called 'the experience sampling method'. Ordinary people were asked to keep an electronic pager for a week which gave out a beeping sound eight times a day. Every time it did so, they
 30 wrote down where they were, what they were doing, how they felt and how much they were concentrating. This system has now been used on more than 10,000 people, and the answers are consistent: as with creative people, ordinary
 35 people are happiest when concentrating hard.



After carrying out 30 years of research and writing 18 books, I believe I have proved that happiness is quite different from what most people imagine. It is not something that can
 40 be bought or collected. People need more than just wealth and comfort in order to lead happy lives. I discovered that people who earn less than £10,000 are not generally as happy as people whose incomes are above that level. This
 45 suggests that there is a minimum amount of money we need to earn to make us happy, but above that dividing line, people's happiness has very little to do with how much poorer or richer they are. Multi-millionaires turn out to be only
 50 slightly happier than other people who are not so rich. What is more, people living below the dividing line and in poverty are often quite happy too.

I found that the most obvious cause of happiness
 55 is intense concentration. This must be the main reason why activities such as music, art, literature, sports and other forms of leisure have survived. In order to concentrate, whether you're reading a poem or building a sandcastle, what
 60 you need is a challenge that matches your ability. The way to remain continually happy, therefore, is to keep finding new opportunities to improve your skills. This may mean learning to do your job better or faster, or doing other more difficult
 65 jobs. As you grow older, you have to find new challenges which are more appropriate to your age. I have spent my life studying happiness and now, as I look back, I wonder if I have achieved it. Overall, I think I have, and my belief that I have
 70 found the keys to its secret has increased my happiness immeasurably.

Adapted from *The Times*

5. Subject the following *True/False* exercise from a student coursebook to the same considerations as the previous exercise type.

7B Natural solutions

Articles

1 Look at the two photos and describe what they show. What do you think the connection is between them?

2 Read a lecture handout about Velcro. Check your answers to 1 and answer the true or false sentences.

a The seeds George de Mestral found had a special quality.	T / F
b Velcro is a natural product.	T / F
c Biomimicry is a complicated idea.	T / F
d Plants and animals can help us solve design problems.	T / F

The invention of Velcro

One day in 1941, Swiss engineer George de Mestral went for a walk with his dog. When he got back, he noticed some plant seeds stuck to the dog's fur. He inspected the seeds more closely to see how they stuck to things so effectively. Using a microscope he saw that each seed had a hook and the hook allowed the seed to stick to anything it touched. De Mestral decided to use the same idea to invent a material which could fasten and attach to things. As a result, Velcro was invented.

The story of Velcro is probably the most famous example of 'biomimicry', the science of copying nature to solve design challenges. The idea behind biomimicry is simple – nature is the best engineer and the plants and animals around us are the perfect models for product designers and scientists to copy.

FOCUS

Articles

The articles *the*, *a* and *an* come at the beginning of a noun phrase. In some cases we do not use an article.

We use *the*:

- when both the speaker/writer and the listener/reader know the thing being referred to
- when there can only be one thing we are referring to
- before a superlative.

Examples Where's Jim? He's in the kitchen.
Neil Armstrong was the first man on the moon.
You're the greatest!

We use *a* and *an*:

- to refer to something for the first time
- to classify or define something
- after *there* is when referring to a single noun.

Examples I saw a man outside the house.
Velcro is a type of material.
There's a spider in the bath.

We don't use an article with plural and uncountable nouns when we are talking about things or people in general.

Example Scientists sometimes copy nature.

(Hughes and Scott-Barrett 2017)

... FURTHER READING

General

Alderson (2000) provides a very full treatment of the testing of reading. Hubley (2012) is a very accessible summary of the issues related to the testing of reading. Weir et al. (2002) describe the development of the specifications of a reading test in China.

Sub-skills

Issues in the testing of reading sub-skills are addressed in Weir et al. (1993), Weir and Porter (1995), Alderson (1990a, 1990b, 1995) and Lumley (1993, 1995). Aryadoust and Zhang (2016) identify two subgroups of readers – one with high lexico-grammatical knowledge, the other with skimming and scanning skills.

Texts in reading tests

Kobayashi (2002) reports on a study which shows how the organisation of a text in a reading test influences the performance of test-takers. Green et al. (2010) use automated textual analysis to compare the appropriacy of texts in tests of academic English.

Multiple choice

Rupp et al. (2006) suggest that multiple choice items prompt test-takers to respond differently from how they would read in a non-testing context. In'nami and Koizumi (2009) compare multiple choice and open-ended formats in reading tests. Shizuka et al. (2006) investigate the merits of reducing the number of multiple choice items in a reading test from four to three.

Other item types

Alderson et al. (2000) explore sequencing as a test technique. Freedle and Kostin (1993) investigate the variables that affect the difficulty of reading items. Trites and McGroarty (2005) report on attempts to design more complex reading tests.

Non-linguistic factors in test performance

Krekeler (2006) investigates the effect of background knowledge on reading test performance. Allan (1992) reports on the development of a scale to measure 'test-wiseness' of people taking reading tests.

12

Testing listening

It may seem rather odd to test listening separately from speaking, since the two skills are typically exercised together in oral interaction. However, there are occasions, such as listening to the radio, podcasts, listening to lectures, online talks and tutorials, or listening to railway station announcements, when no speaking is called for. Also, as far as testing is concerned, there may be situations where the testing of oral ability is considered, for one reason or another, impractical, but where a test of listening is included for its backwash effect on the development of oral skills. Listening may also be tested for diagnostic purposes.

Because it is a receptive skill, the testing of listening parallels in most ways the testing of reading. This chapter will therefore spend little time on issues common to the testing of the two skills and will concentrate more on matters that are particular to listening. The reader who plans to construct a listening test is advised to read both this and the previous chapter.

The special problems in constructing listening tests arise out of the transient nature of the spoken language. Listeners cannot usually move backwards and forwards over what is being said in the way that they can a written text. The one apparent exception to this, when an audio-recording is put at the listener's disposal, does not represent a typical listening task for most people. Ways of dealing with these problems are discussed later in the chapter.

Specifying what the candidate should be able to do

As with the other skills, the specifications for reading tests should say what it is that candidates should be able to do.

Content

Operations

Some operations may be classified as *global*, inasmuch as they depend on an overall grasp of what is listened to. They include the ability to:

- obtain the gist;
- follow an argument;
- recognise the attitude of the speaker.

Other operations may be classified in the same way as were speaking skills in Chapter 10. In writing specifications, it is worth adding to each operation whether what is to be understood is explicitly stated or only implied.

Informational:

- obtain factual information
- follow instructions (including directions)
- understand requests for information
- understand expressions of need
- understand requests for help
- understand requests for permission
- understand apologies
- follow sequence of events (narration)
- recognise and understand opinions
- follow justification of opinions
- understand comparisons
- recognise and understand suggestions
- recognise and understand comments
- recognise and understand excuses
- recognise and understand expressions of preferences
- recognise and understand complaints
- recognise and understand speculation

Interactional:

- understand greetings and introductions
- understand expressions of agreement
- understand expressions of disagreement
- recognise speaker's purpose
- recognise indications of uncertainty
- understand requests for clarification
- recognise requests for clarification
- recognise requests for opinion
- recognise indications of understanding
- recognise indications of failure to understand

- recognise and understand corrections by speaker (of self and others)
- recognise and understand modifications of statements and comments
- recognise speaker's desire that listener indicate understanding
- recognise when speaker justifies or supports statements, etc. of other speaker(s)
- recognise when speaker questions assertions made by other speakers
- recognise attempts to persuade others

It may also be thought worthwhile testing lower-level listening skills in a diagnostic test, since problems with these tend to persist longer than they do in reading. These might include:

- discriminate between vowel phonemes
- discriminate between consonant phonemes
- interpret intonation patterns (recognition of sarcasm, questions in declarative form, etc., interpretation of sentence stress)
- interpret non-verbal information (e.g. facial expressions, gesture)

Texts

For reasons of content validity and backwash, texts should be specified as fully as possible.

Text type might be first specified as monologue, dialogue, or multi-participant, and further specified: conversation, announcement, talk or lecture, instructions, directions, etc.

Text forms include: description, exposition, argumentation, instruction, narration.

Length may be expressed in seconds or minutes. The extent of short utterances or exchanges may be specified in terms of the number of turns taken.

Speed of speech may be expressed as words per minute (wpm) or syllables per second (sps). Reported average speeds for samples of British English are:

	WPM	SPS
Radio monologues	160	4.17
Conversations	210	4.33
Interviews	190	4.17
Lectures to non-native speakers	140	3.17

(Tauroza and Allison 1990)

Dialects may include standard or non-standard varieties.

Accents may be regional or non-regional.

If authenticity is called for, the speech should contain such natural features as assimilation and elision (which tend to increase with speed of delivery) and hesitation phenomena (pauses, fillers, etc.).

Intended audience, style, topics, range of grammar and vocabulary may be indicated.

Increasingly, test developers are incorporating video and other visual information into listening tests. In terms of authenticity this has benefits. Although there are situations, such as listening to the radio, or to airport announcements, where we rely purely on verbal information, these are not the most common. Even traditional 'voice only' phone calls are increasingly being replaced with video calls. In most real-life situations we not only listen, but receive other, non-verbal, information, such as mouth movements, facial expressions, body language or even visual aids. Therefore, tests which contain visual as well as audio information are arguably a better representation of authentic listening. Where visual information is to be included in items, it should of course be included in the test specifications, as in the operations listed above.

Setting criterial levels of performance

The remarks made in the chapter on testing reading apply equally here. If the test is set at an appropriate level, then, as with reading, a near perfect set of responses may be required for a 'pass'. ACTFL, ILR or other scales, including those based on *CEFR*, may be used to validate the criterial levels that are set.

Setting the tasks

Selecting samples of speech (texts)

Passages must be chosen with the test specifications in mind. If we are interested in how candidates can cope with language intended for expert speakers, then ideally we should use samples of authentic speech. These can usually be readily found. Possible sources are podcasts, online lectures, radio, television, teaching materials, and our own recordings of expert speakers. If, on the other hand, we want to know whether candidates can understand language that may be addressed to them as non-expert speakers, suitable examples can be obtained from teaching materials and recordings of expert speakers that we can make ourselves. In some cases the indifferent quality of the recording may necessitate re-recording. It seems to us, although not everyone would agree, that a poor recording introduces difficulties additional to the ones that we want to create, and so reduces the validity of the test. It may also introduce unreliability, since

the performance of individuals may be affected by the recording faults in different degrees from occasion to occasion. If details of what is said on the recording interfere with the writing of good items, testers should feel able to edit the recording, or to make a fresh recording from the amended transcript. In some cases, a recording may be used simply as the basis for a 'live' presentation.

If recordings are made especially for the test, then care must be taken to make them as natural as possible. There is typically a fair amount of redundancy in spoken language: people are likely to paraphrase what they have already said (*'What I mean to say is ...'*), and to remove this redundancy is to make the listening task unnatural. In particular, we should avoid passages originally intended for reading.

Test writers should be wary of trying to create spoken English out of their imagination: it is better to base the passage on a genuine recording, or a transcript of one. If an authentic text is altered, it is wise to check with expert speakers that it still sounds natural. If a recording is made, care should be taken to ensure that it fits with the specifications in terms of speed of delivery, style, etc.

Suitable passages may be of various lengths, depending on what is being tested. A passage lasting ten minutes or more might be needed to test the ability to follow an academic lecture, while twenty seconds could be sufficient to give a set of directions.

Writing items

For extended listening, such as a lecture, a useful first step is to listen to the passage and note down what it is that candidates should be able to get from the passage. We can then attempt to write items that check whether or not they have got what they should be able to get. This note-making procedure will not normally be necessary for shorter passages, which will have been chosen (or constructed) to test particular abilities.

In testing extended listening, it is essential to keep items sufficiently far apart in the passage. If two items are close to each other, candidates may miss the second of them through no fault of their own, and the effect of this on subsequent items can be disastrous, with candidates listening for 'answers' that have already passed. Since a single faulty item can have such an effect, it is particularly important to trial extended listening tests, even if only on colleagues aware of the potential problems.

Candidates should be warned by key words that appear both in the item and in the passage that the information called for is about to be heard. For example, an item may ask about 'the second point that the speaker makes' and candidates will hear 'My second point is ...'. The wording does not have to be identical, but candidates should be given fair warning in the passage. It would be wrong, for instance, to ask about 'what the

speaker regards as her most important point' when the speaker makes the point and only afterwards refers to it as the most important. Less obvious examples should be revealed through trialling.

Other than in exceptional circumstances (such as when the candidates are required to take notes on a lecture without knowing what the items will be, see below), candidates should be given sufficient time at the outset to familiarise themselves with the items. As was suggested for reading in the previous chapter, there seems no sound reason not to write items and accept responses in the native language of the candidates. This will in fact often be what would happen in the real world, when a fellow native speaker asks for information that we have to listen for in the foreign language.

Possible techniques

Multiple choice

The advantages and disadvantages of using multiple choice in extended listening tests are similar to those identified for reading tests in the previous chapter. In addition, however, there is the problem of the candidates having to hold in their heads four or more alternatives while listening to the passage and, after responding to one item, of taking in and retaining the alternatives for the next item. If multiple choice is to be used, then the alternatives must be kept short and simple. The alternatives in the following invented example item are too complex.

Before beginning a journey by car, what is the motorist advised to do?

- a. He should increase the pressure in his tyres to the required level.
- b. He should connect his sat nav and enter his intended destination.
- c. He should make sure that the vehicle is fully roadworthy.
- d. He should ensure that all doors are properly closed, with child locks activated.

Better examples would be:

(Understanding request for help)

I don't suppose you could show me where this goes, could you? Response:

- a. No, I don't suppose so.
- b. Of course I can.
- c. I suppose it won't go.
- d. Not at all.

(Recognising and understanding suggestions)

I've been thinking. Why don't we call Charlie and ask for his opinion?

Response:

- a. Why is this his opinion?
- b. Why do you want to do that?
- c. You think it's his opinion?
- d. Do you think Charlie has called?

Multiple choice can work well for testing lower-level skills, such as phoneme discrimination.

The candidate hears *bat*

and chooses between pat mat fat bat

Short answer

This technique can work well, provided that the question is short and straightforward, and the correct, preferably unique, response is obvious. Below is an example from the *IELTS* test. The candidates hear an extract from a talk given to a group who are going to stay in the UK. Note that the candidates need only give two examples of community groups, with *theatre* provided as an example.

SECTION 2

Questions 11 – 16

Answer the questions below.

Write **NO MORE THAN THREE WORDS AND/OR A NUMBER** for each answer.

What **TWO** factors can make social contact in a foreign country difficult?

- 11
- 12

Which types of community group does the speaker give examples of?

- theatre
- 13
- 14

In which **TWO** places can information about community activities be found?

- 15
- 16

Gap filling

This technique can work well where a short answer question with a unique answer is not possible.

Woman: Do you think you can give me a hand with this?

Man: I'd love to help but I've got to go round to my mother's in a minute.

The woman asks the man if he can _____ her but he has to visit his _____.

Information transfer

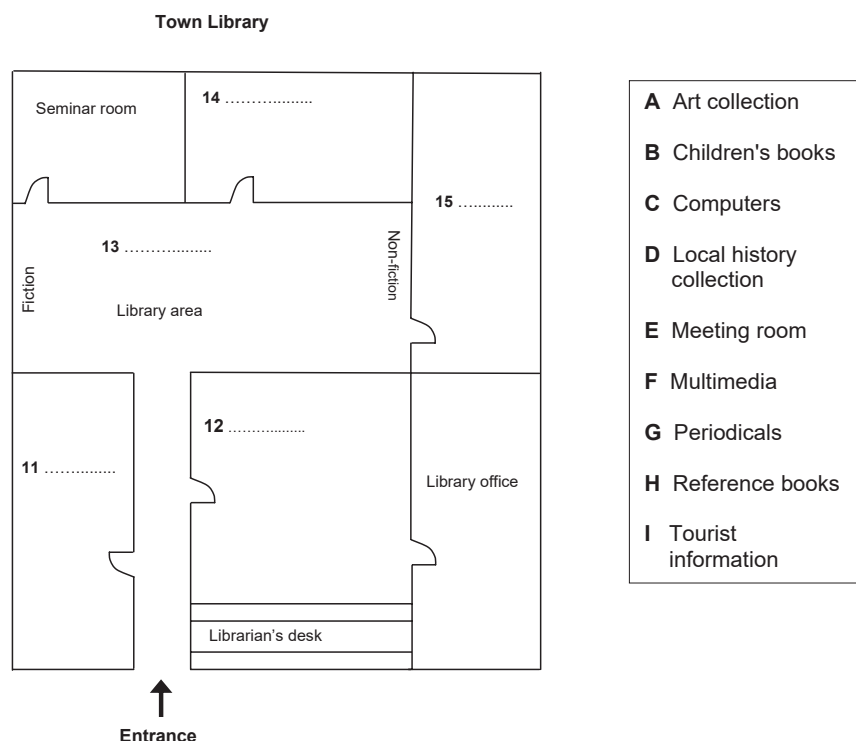
This technique is as useful in testing listening as it is in testing reading, since it makes minimal demands on productive skills. It can involve such activities as the labelling of diagrams or pictures, completing forms, making diary entries, or showing routes on a map. In the following example, which is taken from the *IELTS* exam, candidates label a map while listening to someone describing the layout of a library.

SECTION 2

Questions 11-15

Label the plan below.

Choose **FIVE** answers from the box and write the correct letters **A-I** next to questions 11-15.



You will hear the librarian of a new town library talking to a group of people who are visiting the library.

OK everyone. So here we are at the entrance to the town library. My name is Ann, and I'm the chief librarian here, and you'll usually find me at the desk just by the main entrance here. So I'd like to tell you a bit about the way the library is organised, and what you'll find where ... and you should all have a plan in front of you. Well, as you see my desk is just on your right as you go in, and opposite this the first room on your left has an excellent collection of reference books and is also a place where people can read or study peacefully. Just beyond the librarian's desk on the right is a room where we have up to date periodicals such as newspapers and magazines and this room also has a photocopier in case you want to copy any of the articles. If you carry straight on you'll come into a large room and this is the main library area. There is fiction in the shelves on the left, and non-fiction materials on your right, and on the shelves on the far wall there is an excellent collection of books relating to local history. We're hoping to add a section on local tourist attractions too, later in the year. Through the far door in the library just past the fiction shelves is a seminar room, and that can be booked for meetings or talks, and next door to that is the children's library, which has a good collection of stories and picture books for the under elevens. Then there's a large room to the right of the library area – that's the multimedia collection, where you can borrow videos and DVDs and so on, and we also have CD-Roms you can borrow to use on your computer at home. It was originally the art collection but that's been moved to another building. And that's about it – oh, there's also the Library Office, on the left of the librarian's desk. OK, now does anyone have any questions?

Note taking

Where the ability to take notes while listening to, say, a lecture is in question, this activity can be quite realistically replicated in the testing situation. Candidates take notes during the talk, and only after the talk is finished do they see the items to which they have to respond. When constructing such a test, it is essential to use a passage from which notes can be taken successfully. This will only become clear when the task is first attempted by test writers. We believe it is better to have items (which can be scored easily) rather than attempt to score the notes, which is not a task that is likely to be performed reliably. Items should be written that are perfectly straightforward for someone who has taken appropriate notes. In order to aid authenticity in academic contexts, candidates may be supplied with a copy of the slides used in the lecture. This allows them to make notes on the slides, as they commonly would in their future studies.

It is essential when including note taking as part of a listening test that careful moderation and, if possible, trialling should take place. Otherwise, items are likely to be included that even highly competent speakers of the language do not respond to correctly. It should go without saying that, since this is a testing task which might otherwise be unfamiliar, potential

candidates should be made aware of its existence and, if possible, be provided with practice materials. If this is not done, then the performance of many candidates will lead us to underestimate their ability.

Partial dictation

While dictation may not be a particularly authentic listening activity (although in lectures at university, for instance, there is often a certain amount of dictation), it can be useful as a testing technique. As well as providing a 'rough and ready' measure of listening ability, it can also be used diagnostically to test students' ability to cope with particular difficulties (such as weak forms in English).

Because a traditional dictation is so difficult to score reliably, it is recommended that partial dictation is used, where part of what the candidates hear is already written down for them. It takes the following form:

The candidate sees:

When I _____ someone for the first time,
I _____ them my name. and I always shake their
hand. I think _____ the polite thing to do. I often
_____ nervous when I meet new people so
I _____ play with my hair. I wish I didn't do that.
What do I usually _____ about? The weather and
_____. But I don't talk about _____.
That's _____ rude!

The tester reads:

When I meet someone for the first time, I tell them my name and I
always shake their hand. I think that's the polite thing to do. I often feel
nervous when I meet new people so I sometimes play with my hair. I
wish I didn't do that. What do I usually talk about? The weather and
jobs. But I don't talk about money. That's just rude!

Testers can either write their own passages or they can use authentic transcripts, either from online resources or from student coursebooks, as with the example above. There are advantages to using coursebooks. In addition to the practical benefit of having an audio recording to use, the excerpts from coursebooks will have been written for specific levels of language ability. The possible disadvantage is that some candidates may already be aware of the coursebook. Therefore, we recommend coursebook excerpts only be used in classroom tests. For higher-stakes tests, we suggest it is preferable to use one of the many online resources of authentic listening samples, some of which are listed at the end of this chapter.

Since it is listening that is meant to be tested, correct spelling should probably not be required for a response to be scored as correct. However,

it is not enough for candidates simply to attempt a representation of the sounds that they hear, without making sense of those sounds. To be scored as correct, a response has to provide strong evidence of the candidate's having heard and recognised the missing word, even if they cannot spell it. It has to be admitted that this can cause scoring problems.

The gaps may be longer than one word:

When I meet someone _____, I tell them my name and I always shake their hand.

While this has the advantage of requiring the candidate to do more than listen for a single word, it does make the scoring (even) less straightforward.

Transcription

Candidates may be asked to transcribe numbers or words which are spelled letter by letter. The numbers may make up a telephone number. The letters should make up a name or a word which the candidates should not already be able to spell. The skill that items of this kind test belong directly to the 'real world'. In the trialling of a test we were involved with recently, it was surprising how many teachers of English were unable to perform such tasks satisfactorily. A reliable and, we believe, valid way of scoring transcription is to require the response to an item to be entirely correct for a point to be awarded.

Moderating the items

The moderation of listening items is essential. Ideally it should be carried out using the already prepared recordings or with the item writer reading the text as it is meant to be spoken in the test. The moderators begin by 'taking' the test and then analyse their items and their reactions to them. The moderation checklist given on page 156 for reading items needs only minor modifications in order to be used for moderating listening items.

Presenting the texts (live or recorded?)

The great advantage of using recordings when administering a listening test is that there is uniformity in what is presented to the candidates. This is fine if the recording is to be listened to in a well-maintained language laboratory or in a room with good acoustic qualities and with suitable equipment (the recording should be equally clear in all parts of the room). If these conditions do not obtain, then a live presentation is to be preferred. If presentations are to be live, then the greatest uniformity (and so reliability) will be achieved if there is just a single speaker for each (part of a) test. If the test is being administered at the same time in a number of rooms, more than one speaker will be called for. In either case, a recording should be made of the presentation, with which speakers can be trained, so that the intended emphases, timing, etc. will be observed

with consistency. Needless to say, speakers should have a good command of the language of the test and be generally highly reliable, responsible and trustworthy individuals.

Scoring the listening test

It is probably worth mentioning again that in scoring a test of a receptive skill there is no reason to deduct points for errors of grammar or spelling, provided that it is clear that the correct response was intended.



READER ACTIVITIES

1. a. Choose an online video lecture that would be appropriate for a group of students with whom you are familiar (see end of this chapter for possible resources). Play a five-minute stretch to yourself and take notes. On the basis of the notes, construct eight short-answer items. Ask colleagues to take the test and comment on it. Amend the test as necessary, and administer it without video (audio only) to half of the group of students you had in mind. Analyse the results.
b. Administer the same test to the other half of the group, showing them the video as well as the audio. What differences do you notice between the performance of the two groups of students? Go through the test item by item with the students and ask for their comments. How far, and how well, is each item testing what you thought it would test?
2. Design short items that attempt to discover whether candidates can recognise: sarcasm, surprise, boredom, elation. Try these on colleagues and students.
3. Design a test that requires candidates to draw (or complete) simple pictures. Decide exactly what the test is measuring. Think what other things could be measured using this or similar techniques. Administer the test and see if the students agree with you about what is being measured.



FURTHER READING

General

Buck (2001) is a thorough study of the assessment of listening. Field (2019) evaluates many of the conventions behind listening tests and provides practical ideas for how they might be rethought.

Test methods

Sherman (1997) examines the effects of candidates previewing listening test items. Buck and Tatsuoka (1998) analyse performance on short-answer items. Hale and Courtney (1994) look at the effects of note taking on performance on *TOEFL*® listening items. Note taking is suggested to be a good indicator of listening ability in Song (2012). Shohamy and Inbar (1991) look at the effects of texts and question type. Cai (2013) examines the validity of partial dictation as a test of 'higher order' listening abilities.

The effects of visual information in listening tests are investigated in Ginther (2002), Ockey (2007), Wagner (2010) and Batty (2015).

Test validation

Buck (1991) uses introspection in the validation of a listening test.

Optimising test performance

Arnold (2000) shows how performance on a listening test can be improved by reducing stress in those who take it.

Texts

Freedle and Kostin (1999) investigate the importance of the text in *TOEFL®* minitalk items. Examples of recordings in English that might be used as the basis of listening tests are Crystal and Davy (1975); Hughes et al. (2012), if regional British accents are relevant. Harding (2012) investigates the possibility of bias where accents of speakers in recordings are similar to those of the test-takers' L1. Ockey and Wagner (2018) is a collection of articles on authenticity in the assessment of listening ability.

Online resources

There are countless online resources of authentic spoken English, which testers can use to create tests. What follows is a brief selection of resources that can easily be found using a search engine. The Self-access centre for Language Learning at the University of Reading provides dozens of authentic academic lectures. TED has thousands of talks and lectures on every subject imaginable. Transcripts can be accessed through the TED website. Podcasts are another good way to use authentic listening samples in tests. The BBC website contains hundreds of podcasts in different genres.

13

Testing grammar and vocabulary

Testing grammar

Why test grammar?

Can one justify the separate testing of grammar? There was a time when this would have seemed a very odd question. Control of grammatical structures was seen as the very core of language ability and it would have been unthinkable not to test it. But times have changed. As far as proficiency tests are concerned, there has been a shift towards the view that since it is language skills that are usually of interest, then it is these which should be tested directly, not the abilities that seem to underlie them. For one thing, it is argued, there is more to any skill than the sum of its parts; one cannot accurately predict mastery of the skill by measuring control of what we believe to be the abilities that underlie it. For another, as has been argued earlier in this book, the backwash effect of tests that measure mastery of skills directly may be thought preferable to that of tests that might encourage the learning of grammatical structures in isolation, with no apparent need to use them. Considerations of this kind have resulted in the absence of any grammar component in some well-known proficiency tests.

But probably most proficiency tests that are administered on a large scale still retain a grammar section. One reason for this must be the ease with which large numbers of items can be administered and scored within a short period of time. Related to that, and at least as important, is the question of content validity. If we decide to test writing ability directly, then we are severely limited in the number of topics, styles of writing, and what we earlier referred to as *operations* that we can cover in any one version of the test. We cannot be completely confident that the sample chosen is truly representative of all possibilities. Neither can we be sure, of course, that a (proficiency) grammar test includes a good sample of all possible grammatical elements. But the very fact that there can be so many items does put the grammar test at an advantage.

Even if one has doubts about testing grammar in a proficiency test, there is often good cause to include a grammar component in the achievement, placement and diagnostic tests of teaching institutions. It seems unlikely that there are many institutions, however 'communicative' their approach, that do not teach some grammar in some guise or other. Wherever the

teaching of grammar is thought necessary, then consideration should be given to the advisability of including a grammar component in achievement tests. If this is done, however, it would seem prudent, from the point of view of backwash, not to give such components too much prominence in relation to tests of skills, the development of which will normally constitute the primary objectives of language courses.

Whether or not grammar has an important place in an institution's teaching, it has to be accepted that grammatical ability, or rather the lack of it, sets limits to what can be achieved in the way of skills performance. The successful writing of academic assignments, for example, must depend to some extent on command of more than the most elementary grammatical structures. It would seem to follow from this that in order to place students in the most appropriate class for the development of such skills, knowledge of a student's grammatical ability would be very useful information. There appears to be room for a grammar component in at least some placement tests.

It would be very useful to have diagnostic tests of grammar which could tell us – for individual learners and groups – what gaps exist in their grammatical repertoire. Such tests could inform not only teachers but also learners, so that they could take responsibility for filling the existing gaps themselves. For this reason, it would be important for the tests to be linked in some way or other to learning materials. Unfortunately, as we said in Chapter 3, no fully comprehensive diagnostic test of grammar is yet available. There are, however, partial tests and we point the reader in their direction at the end of this chapter.

Writing specifications

For achievement tests where teaching objectives or the syllabus list the grammatical structures to be taught, specification of content should be quite straightforward. In various parts of the world, there is a growing tendency for coursebooks and syllabuses to be levelled to the *CEFR*. For English in particular, the availability of the *Cambridge Grammar Profile* and the *British Council / EAQUALS* core inventory for General English provide ready-made lists of structure for the different *CEFR* levels.

When there is no such listing it becomes necessary to infer from coursebooks and other teaching materials what structures are being taught. Specifications for a placement test will normally include all of the structures identified in this way, as well as, perhaps, those structures the command of which is taken for granted in even the lowest classes. For proficiency and diagnostic tests, the van Ek and Trim publications referred to in the Further reading section, which are based on a notional-functional approach, are especially useful, as are grammars like the *Cobuild English Usage*.

Sampling

This will reflect an attempt to give the test content validity by selecting widely from the structures specified. It should also take account of what are regarded for one reason or another as the most important structures. It should not deliberately concentrate on the structures that happen to be easiest to test.

Writing items

Whatever techniques are chosen for testing grammar, it is important for the text of the item to be written in grammatically correct and natural language. It is surprising how often this is not the case. Two examples we have to hand from items written by teachers are:

We can't work with this class because there isn't enough silence.
and

I want to see the film. The actors play well.

To avoid unnatural language of this kind, we would recommend using corpus-based examples. One readily available source for English is the *British National Corpus*, which can be accessed free online.

Four techniques are presented for testing grammar: gap filling, paraphrase, completion, and multiple choice. Used with imagination, they should meet just about all our needs. The first three require production on the part of the candidates, while multiple choice, of course, calls only for recognition. This difference may be a factor in choosing one technique rather than another.

Gap filling

Ideally, gap filling items should have just one correct response.

For example: What was most disturbing _____ that for the first time in his life Henry was on his own. [was]

Or: The council must do something to improve transport in the city. _____, they will lose the next election. [Otherwise]
(Sentence linking can be tested extensively using gap filling)

Or: He arrived late, _____ was a surprise. [which]

An item with two possible correct responses may be acceptable if the meaning is the same, whichever is used: Thus:

He displayed the wide, bright smile _____ had charmed so many people before. [which, that]

But an item is probably to be rejected if the different possibilities give different meanings or involve quite different structures, one of which is the one that is supposed to be tested.

Patient: My baby keeps me awake all night. She won't stop crying.

Doctor: _____ let her cry. She'll stop in the end. [Just, I'd, Well, Then, etc.]

This item may be improved by including the words 'Then' and 'just' so that it cannot fill the gap.

Doctor: Then _____ just let her cry. She'll stop in the end.

(But if *you* or *I'd* is thought to be a possible correct response, then the item is still not acceptable.)

It's worth saying here that if contractions like *I'd* are to be allowed in the gaps (and we would recommend this), the possibility should be made very clear to the candidates and at least one example should be given at the beginning of the test.

As was pointed out in Chapter 8, adding to the context can often restrict the number of possible correct responses to a single one. An extension of this is to present a longer passage with several gaps. These may be used to test a set of related structures, such as the articles:

(Candidates are required to write *the*, *a* or *NA* (No Article).)

In England children go to _____ school from Monday to Friday. _____ school that Mary goes to is very small. She walks there each morning with _____ friend. One morning they saw _____ man throwing _____ stones and _____ pieces of wood at _____ dog. _____ dog was afraid of _____ man.

And so on.

The technique can also be used to test a variety of structures, as with the example below, which tests both grammar and vocabulary. (The text is taken from Hughes, *The Pursuit of Truth* (2011))

Yes, I can imagine that, he thought. He sat down _____ front _____ a set of files and began slowly to turn over the sheets of paper that made _____ the first of them. He had hardly begun when Wright arrived. He wished Teague was with him; he didn't fancy _____ this by himself. Still, he would have to.

There can be just a gap, as above, or there can be a prompt for each gap, as in the example below.

Culture shock for international students

Students going to study in another country usually have to make a number of cultural (0) adjustments . They may find it difficult to form (1) with local people and they will certainly have to get used to a (2) of new things including food, the climate and the language. An extra difficulty may be the different (3) which their teachers and tutors have of them in (4) with their home country. They may be (5) for the amount of work they have to do on their own or the fact that their tutors are looking for originality and a capacity for (6) thought rather than an ability to memorise large quantities of information. Equally, they may sometimes be surprised by the (7) of their fellow students who, although usually friendly and (8) , may sometimes seem a little immature. As time passes, international students will find that things become easier and what was unfamiliar to start with will eventually seem normal.

ADJUST
FRIEND
VARY

EXPECT
COMPARE

PREPARE

DEPEND

BEHAVE
WELCOME

Paraphrase

Paraphrase items require the student to write a sentence equivalent in meaning to one that is given. It is helpful to give part of the paraphrase in order to restrict the students to the grammatical structure being tested.

Thus:

1. Testing passive, past continuous form.

When we arrived, a policeman was questioning the bank clerk.

When we arrived, the bank clerk _____

2. Testing present perfect with *for*.

It is six years since I last saw him.

I _____ six years.

The focus in paraphrase items may be grammatical, lexical or both, as can be seen in these examples from the *Cambridge English B2 First Handbook*.

Part 4

For questions 25 – 30, complete the second sentence so that it has a similar meaning to the first sentence, using the word given. **Do not change the word given.** You must use between **two** and **five** words, including the word given. Here is an example (0).

Example:

- 0 A very friendly taxi driver drove us into town.

DRIVEN

We a very friendly taxi driver.

The gap can be filled by the words 'were driven into town by', so you write:

Example: 0 WERE DRIVEN INTO TOWN BY

Write **only** the missing words **IN CAPITAL LETTERS** on the separate answer sheet.

- 25 Joan was in favour of visiting the museum.

IDEA

Joan thought it would be to the museum.

- 26 Arthur has the talent to become a concert pianist.

THAT

Arthur is so could become a concert pianist.

- 27 'Do you know when the match starts, Sally?' asked Mary.

IF

Mary asked Sally time the match started.

- 28 I knocked for ages at Ruth's door but I got no reply.

LONG

I knocking at Ruth's door but I got no reply.

- 29 Everyone says that the band is planning to go on a world tour next year.

SAID

The band planning to go on a world tour next year.

- 30 I'd prefer not to cancel the meeting.

CALL

I'd rather the meeting.

Completion

This technique can be used to test a variety of structures. Note how the context in a passage like the following allows the tester to elicit specific structures, in this case interrogative forms.

In the following conversation, the sentences numbered (1) to (6) have been left incomplete. Complete them suitably. Read the whole conversation before you begin to answer the question. (Michael is attending for interview at a university.)

- Dr Thomson:** Good morning, Michael. Please take a seat. Thank you for applying to our department. (1) Where _____ come from today?
- Michael:** Liverpool.
- Dr Thomson:** A long way! (2) What time _____ your house?
- Michael:** Six o'clock.
- Dr Thomson:** So early? (3) _____ tired?.
- Michael:** No, not really. I slept on the train.
- Dr Thomson:** That's good. (4) Now then, _____ want to study French at university?
- Michael:** Because I have always liked French people. I want to work in France one day.
- Dr Thomson:** That's a good reason. (5) And _____ French at school now?
- Michael:** Yes. The exam is in June.
- Dr Thomson:** Of course.

The telephone rings and Dr Thomson picks up.

Dr Thomson: Oh, hello. I'm conducting an interview at the moment. I'll have to call you back.

He turns to Michael.

Dr Thomson: I'm sorry about that. (6) Now where ?

Oh, yes. I was asking about French at school.

Multiple choice

Reasons for being careful about using multiple choice were given in Chapter 8. There are times, however, when gap filling will not test what we want it to test (at least, in our experience). Here is an example where we want to test epistemic *could*.

If we have the simple sentence:

They left at seven. They _____ be home by now.

There are obviously too many possibilities for the gap (*must, should, may, could, might, will*).

We can add context, having someone reply: *Yes, but we can't count on it, can we?* This removes the possibility of *must* and *will* but leaves the other possibilities.

At this point we would think that we could only test the epistemic use of *could* satisfactorily by resorting to multiple choice.

A: They left at seven. They _____ be home by now.

B: Yes, but we can't count on it, can we?

- a. can b. could c. will d. must

We would also use multiple choice when testing discontinuous elements.

A: Poor man, he _____ at that for days now.

B: Why doesn't he give up?

- a. was working
b. has been working
c. is working
d. had worked

(*Why doesn't he give up?* is added to eliminate the possibility of d being correct, which might just be possible despite the presence of *now*.)

Also, all the above non-multiple choice techniques can be given a multiple choice structure, but the reader who attempts to write such items can often expect to have problems in finding suitable distractors.

Moderation of items is of course essential. The checklist included in Chapter 7 should be helpful in this.

Scoring production grammar tests

Gap filling and multiple choice items should cause no problems for scoring. The important thing when scoring other types of item is to be clear about what each item is testing, and to award points for that only. There may be just one element, such as subject-pronoun-verb inversion, and all available points should be awarded for that; nothing should be deducted for non-grammatical errors, or for errors in elements of grammar which are not being tested by the item. For instance, a candidate should not be penalised for a missing third person -s when the item is testing relative pronouns; *opend* should be accepted for *opened*, without penalty.

If two elements are being tested in an item, then points may be assigned to each of them (for example present perfect form and *since* with past time reference point). Alternatively, it can be stipulated that both elements have to be correct for any points to be awarded, which makes sense in those cases where getting one element wrong means that the student does not have full control of the structure. For items such as these, to ensure scoring is valid and reliable, careful preparation of the scoring key is necessary.

Testing vocabulary

Why test vocabulary?

Similar reasons may be advanced for testing vocabulary in proficiency tests to those used to support the inclusion of a grammar section (though vocabulary has its special sampling problems). However, the arguments for a separate component in other kinds of tests may not have the same strength. One suspects that much less time is devoted to the regular, conscious teaching of vocabulary than to the similar teaching of grammar. If there is little teaching of vocabulary, it may be argued that there is little call for achievement tests of vocabulary. At the same time, it is to be hoped that vocabulary *learning* is taking place. Achievement tests that measure the extent of this learning (and encourage it) perhaps do have a part to play in institutional testing. For those who believe that systematic teaching of vocabulary is desirable, vocabulary achievement tests are appreciated for their backwash effect.

The usefulness (and indeed the feasibility) of a general diagnostic test of vocabulary is not readily apparent. As far as placement tests are concerned, we would not normally require, or expect, a particular set of lexical items to be a prerequisite for a particular language class. All we would be looking for is some general indication of the adequacy of the student's vocabulary. The learning of specific lexical items in class will rarely depend on previous knowledge of other, specified items. One alternative is to use a published test of vocabulary. The other is to construct one's own vocabulary proficiency test.

In this chapter, we will restrict ourselves to the referential meaning of words. However, see the reader activities at the end of this chapter for other aspects of meaning.

Writing specifications

How do we specify the vocabulary for an achievement test? If vocabulary is being consciously taught, then presumably all the items thereby presented to the students should be included in the specifications. To these we can add all the new items that the students have met in other activities (reading, listening, etc.). Words should be grouped according to whether their recognition or their production is required. A subsequent step is to group the items in terms of their relative importance.

We have suggested that a vocabulary placement test will be in essence a proficiency test. The usual way to specify the lexical items that may be tested in a proficiency test is to make reference to one of the published word lists that indicate the frequency with which the words have been found to be used, and, in the case of English, to the *Cambridge English Vocabulary Profile*, or the *Pearson Global Scale of English* (see Further reading).

Sampling

Words can be grouped according to their frequency and usefulness. From each of these groups, items can be taken at random, with more being selected from the groups containing the more frequent and useful words. Some online resources which should help with both sampling and the writing of items will be referred to at the end of this chapter.

Writing items

Testing recognition ability

This is one testing problem for which multiple choice can be recommended without too many reservations. For one thing, distractors are usually readily available. For another, there seems unlikely to be any serious harmful backwash effect, since guessing the meaning of vocabulary items is something that we would probably wish to encourage. However, the writing of successful items is not without its difficulties.

Items may involve a number of different operations on the part of the candidates:

Recognise synonyms

Choose the alternative (a, b, c or d) which is closest in meaning to the word on the left of the page.

gleam a. gather b. shine c. welcome d. clean

The writer of this item has probably chosen the first alternative because of the word *glean*. The fourth may have been chosen because of the similarity of its sound to that of *gleam*. Whether these distractors would work as intended would only be discovered through trialling.

Note that all of the options are words that the candidates are expected to know. If, for example, *welcome* were replaced by *groyne*, most candidates, recognising that it is the meaning of the stem (*gleam*) on which they are being tested, would dismiss *groyne* immediately.

On the other hand, the item could have a common word as the stem with four less frequent words as options:

shine a. malm b. gleam c. loam d. snarl

The drawback to doing this is the problem of what distractors to use. Clearly they should not be too common, otherwise they will not distract. But even if they are not common, if the test-taker knows them, they will not distract. This suggests that the first method is preferable.

Note that in both items it is the word *gleam* that is being tested.

Recognise definitions

loathe means

- a. dislike intensely
- b. become seriously ill
- c. search carefully
- d. look very angry

Note that all of the options are of about the same length. It is said that test-takers who are uncertain of which option is correct will tend to choose the one which is noticeably different from the others. If *dislike intensely* is to be used as the definition, then the distractors should be made to resemble it. In this case the writer has included some notion of intensity in all of the options.

Again the difficult word could be one of the options, although the concern expressed above about this technique applies here too.

One word that means to *dislike intensely* is

- a. growl
- b. screech
- c. sneer
- d. loathe

Recognise appropriate word for context

Context, rather than a definition or a synonym, can be used to test knowledge of a lexical item.

The strong wind _____ the man's efforts to put up the tent.

- a. disabled
- b. hampered
- c. deranged
- d. regaled

Note that the context should not itself contain words that the candidates are unlikely to know.

There are some language testers who insist that it is always better to test vocabulary in context. While we are not averse to including context, as in the above example, we know of no systematic research that has compared test performance on vocabulary items with and without context.

Testing production ability

The testing of vocabulary productively is so difficult that it is practically never attempted in proficiency tests. Information on receptive ability is regarded as sufficient. The suggestions presented below are intended only for possible use in achievement tests.

Pictures

The main difficulty in testing productive lexical ability is the need to limit the candidate to the (usually one) lexical item that we have in mind, while using only simple vocabulary ourselves. One way round this is to use pictures.

Each of the objects drawn below has a letter against it. Write down the names of the objects:



A _____



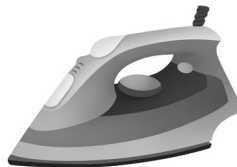
B _____



C _____



D _____



E _____



F _____

This method of testing vocabulary is obviously restricted to concrete nouns that can be unambiguously drawn.

Definitions

This may work for a range of lexical items:

A _____ is a person who looks after our teeth.

_____ is frozen water.

_____ is the second month of the year.

But not all items can be identified uniquely from a definition: any definition of, say, *feeble* would be unlikely to exclude all of its synonyms. Nor can all words be defined entirely in words more common or simpler than themselves.

Gap filling

This can take the form of one or more sentences with a single word missing.

Because of the snow, the football match was _____ until the following week.

I _____ to have to tell you this, Mrs Jones, but your husband has had an accident.

Too often there is an alternative word to the one we have in mind. Indeed the second item above has at least two acceptable responses (which was not intended when it was written!). This problem can be solved by giving the first letter of the word (possibly more) and even an indication of the number of letters.

I r _____ to have to tell you ...

or I r _ _ _ _ _ to have to tell you ...

Again, moderation of items is necessary and the checklist in Chapter 7 can be used, possibly with minor modifications.

Readers will notice that we are not recommending one item format above all others. Rather, we believe that item writers must decide on which format is most appropriate for the specific vocabulary item being tested. For example, picture matching may work well for a concrete noun such as *shoes*, while gap filling, where more context can be provided, would be better for an adverb such as *usually*.

Postscript

This chapter should end with a reminder that while grammar and vocabulary contribute to communicative skills, they are rarely to be regarded as ends in themselves. It is essential that tests should not accord them too much importance, and so create a backwash effect that undermines the achievement of the objectives of teaching and learning where these are communicative in nature.



READER ACTIVITIES

1. Construct items to test the following:

- Conditional: *If ... had would have ...*
- Comparison of equality.
- Relative pronoun *whose*.
- Past continuous: *... was -ing, when ...*

Which of the techniques suggested in the chapter suits each structure best? Can you say why?

2. Can you see anything wrong with the following multiple choice items taken from tests written by teachers (use the checklist given as Table 2 in Chapter 7). If so, what? Try to improve them.

- a. I said to my friend '_____ be stupid.'

Isn't Aren't Didn't Don't be

- b. What _____ you do, if your car broke down?

must did shall

- c. You are too thin. You should eat _____

many more a few

- d. I'm sorry that the child saw the accident.

– I don't think it matters. He soon _____ it.

is forgetting forgets will forget will be forgetting

- e. People _____ in their reaction to the same stimulus.

replace vary upset very

3. Produce three vocabulary tests by writing three items for each of the following words. One set of items should be multiple choice *without* context; one set should be multiple choice *with* context; the third set should be gap filling. Give each test to a different (but comparable) group of students. Compare performance on items testing the same word. Can differences of performance be attributed to a difference in technique?

beard	sigh	bench	deaf	genial
tickle	weep	greedy	mellow	callow

(If the words are inappropriate for your students, replace them with others.)

4. Connotation and collocation are notoriously difficult to test but they could well form part of the non-testing assessment of vocabulary (see Chapter 16). How would you assess a student's control of connotation and collocation? Give two examples of each.
5. Look at the paraphrase items from the *Cambridge English B2 First Handbook* on pages 181–182. For each item, identify whether it is testing grammar, vocabulary or both. Compare with a colleague.



FURTHER READING

Testing vocabulary

Dávid (2007) evaluates a modified type of multiple choice item to assess grammatical knowledge. Laufer and Goldstein (2004) describe the development of a computerised test which assesses a candidate's vocabulary knowledge, both in terms of how many words they know, and how well they know the words. Qian and Schedl (2004) investigate an in-depth vocabulary measure as a predictor for test-takers' reading performance in the *TOEFL®*. Read (2000) is a thorough study of vocabulary assessment (going beyond testing). It includes methods of assessing both size (breadth) and quality (depth) of knowledge. Read and Chapelle (2001) propose a framework for vocabulary assessment. Read (2007) discusses the usefulness of various corpora in relation to the assessment of vocabulary. Alderson (2005) and van Ek and Trim (2001a and b) below also relate to vocabulary.

Testing grammar

Alderson (2005) gives an account of the development of *DIALANG*. Chapelle et al. (2010) is a promising, and very interesting, investigation into a computer-delivered test of productive grammatical ability based on Second Language Acquisition findings.

Rimmer (2006) investigates grammatical complexity and its role in describing grammatical competence. van Ek and Trim (2001a, b and c) provide a highly detailed taxonomy of notions and functions and their grammatical and lexical realisations.

Online resources

Since the last edition of this book, there has been a rapid increase in the number of helpful online resources, particularly for English, some of which are listed below.

English Vocabulary Profile and *English Grammar Profile* are two online tools which help to show what the different levels of the *CEFR* mean in relation to vocabulary and grammar.

The *British Council / EAQUALS Core Inventory for General English* also lists structures for the various *CEFR* levels.

The *Pearson Global Scale of English* is another useful resource for the development of English language tests.

The *Oxford 3000™* is a list of the 3,000 words considered to be the most important words to learn in the English language.

The *British National Corpus* contains a huge number of samples of spoken and written British English. It can be accessed through various interfaces.

Designed by Dr Averil Coxhead, the *Academic Word List* contains over 500 word families which frequently appear within academic texts.

Using data from the *Corpus of Contemporary American English*, wordandphrase is a very user-friendly website which gives information on the words and phrases within any text you submit.

TEDDCLOG is an automatic gap-fill generator.

There are also various 'diagnostic' grammar and vocabulary tests online, which can be found by typing the relevant keywords into a search engine. These tests are not comprehensive but readers may find them useful and interesting nonetheless.

14

Testing overall ability

The previous five chapters have given advice on the testing of different abilities. The assumption has been that we need to obtain separate information on each of these abilities. There are circumstances, however, when we do not need such detailed information, when an estimate of candidates' overall ability is enough.

We will begin with a discussion of the notion of overall ability, then identify ways in which it may be measured, before going on to outline the circumstances in which tests of overall ability may reasonably be used.

Overall ability as a concept

The notion of overall ability is directly related to the common-sense idea that someone can be good (quite good, or poor) at a second or foreign language. It makes sense to say that someone is good at a language because performance in one skill is usually a reasonable predictor of performance in another. If we hear someone speaking a language fluently and correctly, we can predict that they will also write the language well. On some occasions, of course, we will be wrong in our prediction (particularly where teaching has favoured one skill over another), but usually we will be right. This is hardly surprising, since, despite their differences, speaking and writing share a great many features, most obviously elements of grammar and vocabulary. It is essentially this sharing of features that allows us to measure overall ability. It is worth pointing out that major tests such as the *Cambridge English C2 Proficiency* implicitly acknowledge the concept of overall ability by awarding a pass based on a candidate's performance on tasks requiring a variety of language skills.

Measuring overall ability

Most techniques for measuring overall ability are based on the idea of *reduced redundancy*. When we listen to someone or read something, there is more information available to us than we actually need in order to interpret what is said or written. There is redundancy. Expert speakers of a language can cope well when this redundancy is reduced. They can, for example, understand what someone is saying even though there are noises in the environment that prevent them from hearing every sound that is made. Similarly, they can make out the meaning of the text of a newspaper that has been left outside in the rain, causing the print to become blurred.

Because non-expert speakers generally find it more difficult to cope with reduced redundancy, the deliberate reduction of redundancy has been used as a means of estimating foreign language ability. Learners' overall ability has been estimated by measuring how well they can restore a reduced text to its original form.

Varieties of cloze procedure

Traditional cloze

In its original form, the cloze procedure reduces redundancy by deleting a number of words in a passage, leaving blanks, and requiring the person taking the test to attempt to replace the original words. After a short unmutated 'lead-in', it is usually about every seventh word that is deleted. The following example, which the reader might wish to attempt, was used in research into cloze in the United States (put only one word in each space). The answers are at the end of this chapter.

What is a college?

Confusion exists concerning the real purposes, aims, and goals of a college. What are these? What should a college be?

Some believe that the chief function 1. _____ even a liberal arts college is 2. _____ vocational one. I feel that the 3. _____ function of a college, while important, 4. _____ nonetheless secondary. Others profess that the 5. _____ purpose of a college is to 6. _____ paragons of moral, mental, and spiritual 7. _____ Bernard McFaddens with halos. If they 8. _____ that the college should include students 9. _____ the highest moral, ethical, and religious 10. _____ by precept and example, I 11. _____ willing to accept the thesis. I 12. _____ in attention to both social amenities 13. _____ regulations, but I prefer to see 14. _____ colleges get down to more basic 15. _____ and ethical considerations instead of standing *in loco parentis* 16. _____ four years when 17. _____ student is attempting in his youthful 18. _____ awkward ways, to grow up. It 19. _____ been said that it was not 20. _____ duty to prolong adolescences. We are 21. _____ adept at it.

There are those 22. _____ maintain that the chief purpose of 23. _____ college is to develop "responsible citizens".

24. _____ is good if responsible citizenship is
 25. _____ by-product of all the factors which
 26. _____ to make up a college education
 27. _____ life itself. The difficulty arises from
 28. _____ confusion about the meaning of
 responsible 29. _____. I know of one college
 which 30. _____ mainly to produce, in a kind
 31. _____ academic assembly line, outstanding
 exponents of 32. _____ system of free enterprise.

Likewise, I 33. _____ to praise the kind of
 education 34. _____ extols one kind of economic
 system 35. _____ the exclusion of the good
 portions 36. _____ other kinds of economic
 systems. It 37. _____ to me, therefore, that a
 college 38. _____ represent a combination of all
 39. _____ above aims, and should be something
 40. _____ besides – first and foremost – an
 educational 41. _____, the center of which is
 the 42. _____ exchange between teachers and
 students.

I 43. _____ read entirely too many statements
 such 44. _____ this one on admissions application
 papers: "45. _____ want a college education
 because I 46. _____ that this will help to support
 47. _____ and my family. "I suspect that
 48. _____ job as a bricklayer would help this
 49. _____ to support himself and his family
 50. _____ better than a college education.

(Oller and Conrad 1971)

Some of the blanks you will have completed with confidence and ease. Others, even if you are an expert speaker of English, you will have found difficult, perhaps impossible. In some cases, you may have supplied a word which, although different from the original, you think is just as good or even better. All of these possible outcomes are discussed in the following pages.

Selected deletion cloze

Even though scores on cloze tests of this kind have tended to correlate highly with scores on longer, more complex and well-established tests, there seems to be general agreement now that the cloze procedure cannot be depended upon automatically to produce reliable and useful tests. There is need for careful selection of texts and some pre-testing. The fact that deletion of every *n*th word almost always produces problematic

items (for example, impossible to predict the missing word), points to the advisability of a careful selection of words to delete, from the outset. The following is an in-house cloze passage, for students at university entrance level, in which this has been done. Again the reader is invited to try to complete the gaps. Again, the answers are at the end of the chapter.

Choose the best word to fill each of the numbered blanks in the passage below. Write your answers in the space provided in the right hand margin. Write only ONE word for each blank.

Ecology

Water, soil and the earth's green mantle of plants make up the world that supports the animal life of the earth. Although modern man seldom remembers the fact, he could not exist without the plants that harness the sun's energy and manufacture the basic food-stuffs he depends (1) _____ for life. Our attitude (2) _____ plants is a singularly narrow (3) _____. If we see any immediate utility in (4) _____ plant we foster it. (5) _____ for any reason we find its presence undesirable, (6) _____ merely a matter of indifference, we may condemn (7) _____ to destruction. Besides the various plants (8) _____ are poisonous to man or to (9) _____ livestock, or crowd out food plants, many are marked (10) _____ destruction merely because, according to our narrow view, they happen to (11) _____ in the wrong place at the (12) _____ time. Many others are destroyed merely (13) _____ they happen to be associates of the unwanted plants.

The earth's vegetation is (14) _____ of a web of life in which there are intimate and essential relations between plants and the earth, between plants and (15) _____ plants, between plants and animals. Sometimes we have no (16) _____ but to disturb (17) _____ relationships, but we should (18) _____ so thoughtfully, with full awareness that (19) _____ we do may (20) _____ consequences remote in time and place.

- (1) _____
- (2) _____
- (3) _____
- (4) _____
- (5) _____
- (6) _____
- (7) _____
- (8) _____
- (9) _____
- (10) _____
- (11) _____
- (12) _____
- (13) _____
- (14) _____
- (15) _____
- (16) _____
- (17) _____
- (18) _____
- (19) _____
- (20) _____

The deletions in the above passage were chosen to provide 'interesting' items. Most of them we might be inclined to regard as testing 'grammar', but to respond to them successfully more than grammatical ability is needed; processing of various features of context is usually necessary. Another feature is that native speakers of the same general academic ability as the students for whom the test was intended could be expected to provide acceptable responses to all of the items. The acceptable responses are themselves limited in number. If cloze is to be used to

measure overall ability, it is this kind which we would recommend. General advice on the construction of such tests is given below.

Multiple choice cloze

Cloze passages can be made multiple choice, making scoring easy and reliable. The *Cambridge English C2 Proficiency* exam includes such a passage in its Reading and Use of English section.

C2 Proficiency (CPE) Practice Test

CPE Reading and Use of English Part 1: Multiple Choice Cloze

For Questions 1-8, read the text below and decide which answer (A, B, C or D) best fits each gap.

Planetary Artistry

By Johanna Kieniewicz

For me, the highlight of this past week's science news was the images (1) back from the Curiosity rover, providing (2) geologic evidence that water flowed on Mars. Of course, this wasn't exactly a surprise; for decades, planetary scientists have suggested the channel networks visible in spacecraft imagery couldn't have been made by anything else. The evidence has been (3) as well, as various clay minerals and iron oxides have been identified through hyperspectral imagery.

Nonetheless, I suspect that the image of definitely water-lain (4) made the heart of more than one geologist (5) a beat. Ground truth. You could argue that the scientific exploration of the extra-terrestrial is, at least (6) part, a search for meaning: to position us within a larger cosmology. But our fascination with, and connection to, what we see in the night sky comes not just through science, but also through art. So it should come as no surprise that scientific images of planetary surfaces have (7) inspiration to a range of artists from Galileo - whose first sketches of the moon through a telescope are (8) beautiful - to Barbara Hepworth - whose interpretations of the lunar surface are far less literal.

Questions

Gap 1

- A. ? thrown
- B. ? shot
- C. ? beamed
- D. ? fired

Gap 2

- A. ? final
- B. ? conclusive
- C. ? proven
- D. ? guaranteed

Gap 3

- A. ? swelling
- B. ? expanding
- C. ? increasing
- D. ? mounting

Gap 5

- A. ? slip
- B. ? lose
- C. ? skip
- D. ? jump

Gap 7

- A. ? offered
- B. ? provided
- C. ? given
- D. ? made

Gap 4

- A. ? sediments
- B. ? dross
- C. ? grounds
- D. ? matter

Gap 6

- A. ? with
- B. ? in
- C. ? at
- D. ? for

Gap 8

- A. ? totally
- B. ? doubtlessly
- C. ? surely
- D. ? truly

Our earlier warnings about the difficulty of writing good multiple choice items apply here too. For teacher-made tests we would recommend requiring candidates to supply their own words.

Conversational cloze

The two passages used to create cloze tests above are both quite formal prose. If we want our measure of overall ability to reflect (and hopefully predict) oral as well as written ability, we can use passages which represent spoken language. The next passage is based on a tape-recording of a conversation. As this type of material is very culturally bound, probably only a non-expert speaker who has been in Britain for some time could understand it fully. It is a good example of informal family conversation, where sentences are left unfinished and topics run into each other. (Again the reader is invited to attempt to predict the missing words. Note that things like *John's*, *I'm*, etc. count as one word. Only one word per space.)

Family reunion

Mother: I love that dress, Mum.
 Grandmother: Oh, it's M and S.
 Mother: Is it?
 Grandmother: Yes, five pounds.
 Mother: My goodness, it's not, Mum.

- Grandmother: But it's made of that T-shirt stuff, so I don't think it'll wash very (1), you know, they go all ...
- Mother: sort (2) ... I know the kind, yes ...
- Grandmother: Yes.
- Mother: I've got some T-shirts of that, and (3) shrink upwards and go wide ...
- Grandmother: I know, so ...
- Mother: It's a super colour. It (4) a terribly expensive one, doesn't it? (5) you think so when you saw (6)?
- Grandmother: Well, I always know in Marks. (7) just go in there and ... and (8) it's not there I don't buy it. I know I won't like anything else. I got about three from there ... four from there. Only I wait about ...
- Girl: Mummy, can I have a sweetie?
- Mother: What, love?
- Grandmother: Do you know what those are called? ... Oh, I used to love them (9) I was a little girl. Liquorice comfits. Do you like liquorice? Does she?
- Mother: (10) think she quite likes it. Do (11)? We've got some liquorice allsorts actually (12) the journey.
- Grandmother: Oh yes.
- Mother: And I said she could have one after.
- Grandmother: Oh, I'm going to have one. No, I'm (13). No, it'd make me fat, dear.
- Mother: Listen. Do you want some stew? It's hot now.
- Grandmother: No, no, darling. I don't want anything.
- Mother: Don't you want any? Because (14) just put it on the table.
- Grandmother: I've got my Limmits.
- Mother: Are you going (15) eat them now with us?
- Grandmother: Yes. (16) you going to have yours ... yours now?
- Mother: Well, I've just put mine on the plate, but Arth says he doesn't (17) any now.
- Grandmother: Oh yes, go on.
- Mother: So ... so he's going to come down later ...
- Grandmother: What are (18) going to eat? ... Oh, I like (19). Is that a thing that ...
- Mother: ... you gave me, but I altered it.
- Grandmother: Did (20) shorten it?
- Mother: I took the frill (21).
- Grandmother: I thought it looked ...
- Mother: I altered (22) straps and I had to ...
- Girl: That's (23) you gave me, Granny....
- Granny, I'm (24) big for that ...
- Mother: And so is Jake. It's for a doll ... Do you remember that?
- Grandmother: No.
- Mother: Oh, Mum, you're awful. (25) made it.

This 'conversational cloze' passage turned out to be a reasonable predictor of the oral ability of overseas students (as rated by their language teachers) who had already been in Britain for some time. It suggests that we should base cloze tests on passages that reflect the kind of language that is relevant for the overall ability we are interested in.



ADVICE ON CREATING CLOZE TYPE PASSAGES

1. The chosen passages should be at a level of difficulty appropriate to the people who are to take the test. If there is doubt about the level, a range of passages should be selected for trialling. Indeed, it is always advisable to trial a number of passages, as their behaviour is not always predictable.
2. The text should be of a style appropriate to the kind of language ability being tested.
3. After a couple of sentences of uninterrupted text, deletions should be made at about every eighth or tenth word (the so-called pseudo-random method of deletion). Individual deletions can then be moved a word or two to left or right, to avoid problems or to create interesting 'items'. One may deliberately make gaps that can only be filled by reference to the extended context.
4. The passage should then be tried out on a good number of comparable expert speakers and the range of acceptable responses determined.
5. Clear instructions should be devised. In particular, it should be made clear what is to be regarded as a word (with examples of *isn't*, etc., where appropriate). Students should be assured that no one can possibly replace all the original words exactly. They should be encouraged to begin by reading the passage right through to get an idea of what is being conveyed (the correct responses early in the passage may be determined by later content).
6. The layout of the second test in the chapter (Ecology) facilitates scoring. Scorers are given a card with the acceptable responses written in such a way as to lie opposite the candidates' responses.
7. Anyone who is to take a cloze test should have had several opportunities to become familiar with the technique. The more practice they have had, the more likely it is that their scores will represent their true ability in the language.
8. Cloze test scores are not directly interpretable. In order to be able to interpret them we need to have some other measure against which they can be validated.

The C-Test

The C-Test is really a variety of cloze, which its originators claim is superior to the kind of cloze described above. Instead of whole words, it is the second half of every second word that is deleted. The following example is one of many available to take online at the Universitat Autònoma de Barcelona (UAB) website.

Pigeons

A new law which came into force last Monday bans the feeding of pigeons in London's Trafalgar Square. Anyone cau() giving fo() to th() will fa() a fine o() up t() £500. Si() 2002, diff() ways o() frightening t() pigeons aw() have be() tried b() none ha() worked. T() London Ci() Council h() spent £25m upgr() the squ(). One Counc() said "t() improvements wo() not wo() if t() square w() still infested with pigeons". However, pigeon supporters plan to ignore the new law and will continue to feed the birds.

The correct responses are to be found at the end of the chapter.

The supposed advantages of the C-Test over the more traditional cloze procedure are that only exact scoring is necessary (expert speakers effectively scoring 100 percent) and that shorter (and so more) passages are possible. This last point means that a wider range of topics, styles and levels of ability is possible. The deletion of elements less than the word is also said to result in a representative sample of parts of speech being so affected. By comparison with cloze, a C-Test of 100 items takes little space and not nearly so much time to complete (candidates do not have to read as much text).

Possible disadvantages relate to the puzzle-like nature of the task. It is harder to read than a cloze passage, and correct responses can often be found in the surrounding text. Thus, the candidate who adopts the right puzzle-solving strategy may be at an advantage over a candidate of similar foreign language ability. However, research would seem to indicate that the C-Test functions well as a rough measure of overall ability in a foreign language. The advice given above about the development of cloze tests applies equally to the C-Test.

Dictation

In the 1960s it was usual, at least in some parts of the world, to decry dictation testing as hopelessly misguided. After all, since the order of words was given, it did not test word order; since the words themselves were given, it did not test vocabulary; since it was possible to identify words from the context, it did not test aural perception. While it might test punctuation and spelling, there were clearly more economical ways of doing this.

This orthodoxy has been challenged, with research showing high correlations between scores on dictation tests and scores on much longer and more complex tests as was the case with cloze. Examination of

performance on dictation tests made it clear that words and word order were not really given; the candidate heard only a stream of sound which had to be decoded into a succession of words, stored and recreated on paper. The ability to identify words from context was now seen as a very desirable ability, one that distinguished between learners at different levels.

Dictation forms part of the Listening section of the *Pearson Test of English (PTE)*. Unrelated sentences are read one at a time, once only, and the candidates have to write what they hear. An example provided on the website for practice is: 'I went to the University of Bristol where I studied chemistry' (read in 3 seconds).

Dictation tests give results similar to those obtained from cloze tests. In predicting overall ability they have the advantage of involving listening ability. That is probably the only advantage. Certainly they are as easy to create. They are relatively easy to administer, though not as easy as the paper-and-pencil cloze. But they are certainly not easy to score. Initial proponents of dictation recommended that a candidate's score should be the number of words appearing in their original sequence (misspelled words being regarded as correct as long as no phonological rule is broken). This works quite well when performance is reasonably accurate, but is still time-consuming. With poorer students, scoring becomes very tedious.

Because of this scoring problem, partial dictation (see pages 172–173) may be considered as an alternative. In this, part of what is dictated is already printed on the candidate's answer sheet. The candidate has simply to fill in the gaps. It is then clear just where the candidate is up to, and scoring is likely to be more reliable.

When using dictation, the same considerations should guide the choice of passages as with the cloze procedure. The passage has then to be broken down into stretches that will be spoken without a break. These should be fairly long, beyond rote memory, so that the candidates will have to decode, store and then re-encode what they hear (this was a feature of the dictations used in the research referred to above). It is usual, when administering the dictation, to begin by reading the entire passage straight through. Then the stretches are read out, not too slowly, one after the other with enough time for the candidates to write down what they have heard.

Elicited imitation

Elicited imitation is normally carried out on a one-to-one basis. It requires a candidate to repeat a series of spoken sentences of increasing length and complexity. Scoring systems vary but the most straightforward is dichotomous: 1 for a completely accurate imitation; 0 for anything else. As a measure of overall ability, the attraction of this technique is that it involves the candidate in speaking, as well as processing

linguistic input. However, unless the testing is carried out entirely on the computer, perhaps using a computer adaptive testing program, it is quite uneconomical¹.

Using measures of overall ability

When would we want to use these measures of overall ability? We believe that they may be used successfully for screening, placement, and as part of a larger test.

Screening

In screening, we eliminate candidates who could not possibly be successful on a longer test which takes time to administer and score. Only those candidates who pass the screening test take the longer one. The measures of overall ability which we have identified above can serve as the basis of screening tests.

Placement tests need not always give detailed information on each candidate. This will often be the case in language schools, where a test of overall ability (preferably supplemented by a brief interview) can be sufficient to place students in appropriate classes, with the knowledge that any errors can easily be corrected early in the course. Where there is a wide range of ability amongst students accepted for courses, tests can be constructed at different levels, the students taking the easiest first. Scoring can begin at the lowest level and stop once it is clear that a student has reached a level at which he or she cannot cope².

Component of larger test

It is not uncommon for the techniques we have described above to be included in larger tests. While the rationale for this is not always made clear, we can see three benefits.

The first concerns reliability. Since the techniques properly used are reliable in nature, their inclusion will tend to increase the reliability of the whole test.

The second concerns validity. By allowing candidates to demonstrate their ability in another, additional way, the effect of any method bias is potentially reduced, and is consistent with the increasing demand for multiple measures in assessment.

1. To prevent candidates using purely rote memory, without the need to process what they hear, their imitation may be delayed until they have performed a simple task (simple arithmetic, adding small numbers to each other, has been used; but on a computer other tasks may easily be devised).
2. We recognise that many language schools are in a position to create more elaborate placement tests or to use commercially available tests. But we also know that this is not the case throughout the world and for all languages that are taught.

The third relates to validation. Correlations between scores on such items and on the other components of a test may offer insights into the functioning of the other components and of the test as a whole.

We have already pointed to the use of multiple choice cloze in the *Cambridge English C2 Proficiency*. Other examples are dictation and elicited imitation as parts of the *Pearson Test of English*.



READER ACTIVITIES

1. Complete the three cloze passages in the chapter. Say what you think each item is testing. How much context do you need to arrive at each correct response?

If there are items for which you cannot provide a satisfactory response, can you explain why?

Identify items for which there seem to be a number of possible acceptable responses. Can you think of responses that are on the borderline of acceptability? Can you say why they are on the borderline?

2. Choose a passage that is at the right level and on an appropriate topic for a group of students with whom you are familiar. Use it to create tests by:
 - deleting every seventh word after a lead-in;
 - doing the same, only starting three words after the first deleted word of the first version.

Compare the two versions. Are they equivalent?

Now use one of them to create a cloze test of the kind recommended.

Make a C-Test based on the same passage. Make a partial dictation of it too. How do all of them compare?

If possible administer them to the group of students you had in mind, and compare the results (with each other and with your knowledge of the students).



FURTHER READING

Cloze

For all issues discussed in this chapter, including dictation, the most accessible source is Oller (1979). The research in which the first cloze passage in the chapter was used is described in Oller and Conrad (1971). Chapelle and Abraham (1990) used one passage but different methods of cloze deletion (including C-Test) and obtained different results with the different methods. Brown (1993) examines the characteristics of 'natural' cloze tests and argues for rational deletion. Farhady and Keramati (1996) propose a 'text-driven' procedure for deleting words in cloze passages. Storey (1997) investigates the processes that candidates go through when taking cloze tests. Hughes (1981) is an account of the research into conversational cloze.

C-Test

Klein-Braley and Raatz (1984) and Klein-Braley (1985) outline the development of the C-Test. Klein-Braley (1997) is a more recent appraisal of the technique. Jafarpur (1995) reports rather negative results for C-Tests he administered. Lee-Ellis (2009) demonstrates that C-Tests can be constructed successfully for learners of Korean. Harsch and Hartig (2016) present evidence for the use of the C-Test for placement and screening purposes.

Drackert and Timukova (2020) offer insights into what the C-Test measures and make suggestions for future research.

Dictation

Lado (1961) provides a critique of dictation as a testing technique, while Lado (1986) carried out further research using a passage employed by Oller and Conrad, to cast doubt on their claims.

Elicited imitation

Yan et al. (2016) review research into elicited imitation over a period of 40 years and argue for the validity of the technique as a measure of second language proficiency.

Answers

What is a college? The words deleted from the passage are as follows:

1. of; 2. a; 3. vocational; 4. is; 5. chief; 6. produce; 7. stamina; 8. mean; 9. with; 10. standards; 11. am; 12. believe; 13. and; 14. our; 15. moral; 16. for; 17. the; 18. and; 19. has; 20. our; 21. singularly; 22. who; 23. a; 24. This; 25. a; 26. go; 27. and; 28. a; 29. citizenship; 30. aims; 31. of; 32. our; 33. hesitate; 34. which; 35. to; 36. of; 37. seems; 38. should; 39. the; 40. else; 41. experience; 42. intellectual; 43. have; 44. as; 45. I; 46. feel; 47. me; 48. a; 49. student; 50. much.

Ecology. The words deleted from the passage are as follows: 1. on; 2. to; 3. one; 4. a; 5. If; 6. or; 7. it; 8. which/that; 9. his; 10. for; 11. be; 12. wrong; 13. because; 14. part; 15. other; 16. choice/option; 17. these; 18. do; 19. what; 20. have.

Family reunion. Acceptable responses: 1. well; 2. of; 3. they; 4. looks, seems; 5. Did, Didn't; 6. it; 7. I; 8. if; 9. when; 10. I; 11. you; 12. for; 13. not; 14. I've; 15. to; 16. Are; 17. want; 18. you; 19. that; 20. you; 21. off; 22. the; 23. what, one; 24. too; 25. You.

Answers to C-Test

ght/od/em/ce/f/o/nce/erent/f/he/ay/en/ut/ve/he/ty/as/ading/are/illor/he/uld/
rk/he/as

15 Tests for young learners

Over the past few decades, the learning of foreign languages at primary school has become increasingly common in many parts of the world. This chapter begins by suggesting a general approach to tests for young learners. It then goes on to consider the particular requirements of such tests. Finally it recommends suitable testing techniques.

General approach

One might first ask why we should test young learners at all. This is a good question. Not everyone does it. In Norway, for example, where the learning of English appears to be highly successful, children up to the age of thirteen are not formally tested in the subject. One answer to the question might be that we want to be sure that the teaching programme is effective, that the children are really benefiting from the chance to learn a language at an early age. But this invites a further question: why is testing rather than a less formal means of assessment necessary? The answer we gave in Chapter 1 was that there is a need for a common yardstick, which tests give, in order to make meaningful comparisons. We have to confess, however, that we feel uneasy at the thought of the damage to children's learning, and their attitude to learning, that might be done by insensitive, inappropriate testing. This uneasiness is not lessened by the knowledge that the aims of early language teaching typically include the development of positive attitudes to language learning and to language.

But young learners *are* tested and, as reported in Rixon's 2013 survey of English language teaching at primary school in 64 countries, the tests they take are often high-stakes. In some parts of the world (for example, in Bahrain and Cameroon) children take English tests at the end of primary school in order to determine which secondary school they will move on to. In some cases, the results of these tests are even used in deciding whether a child is ready to start secondary school at all. Rixon also reports that these tests are often created within the schools themselves, rather than by external professional test designers.

On a more positive note, it seems to us that if young children are going to be tested, such testing provides an opportunity to develop positive attitudes towards assessment, and to help them recognise the value of assessment. In order to take advantage of this opportunity, we would make a number of general recommendations which, together, amount to an approach to such testing.

The first general recommendation is that a special effort be made to make testing an integral part of assessment, and assessment an integral part of the teaching programme. All three elements should be consistent with each other in terms of learning objectives and, as far as possible, with the kinds of tasks which the children are expected to perform. Testing will not then be seen as something separate from learning, as a trial that has to be endured. Clearly these principles apply at later stages of education as well, but we consider it particularly important that a healthy attitude towards testing be instilled at an early age. Learners should see testing as a normal part of learning and in no way threatening.

The second general recommendation is that the children be provided with constant feedback on their test performance. Feedback has already been discussed in an earlier chapter but its provision is especially important for children. In particular:

- Feedback should be immediate and positive. By being immediate, its effect will be maximised. By telling children not only what their weaknesses are, but also what they have done well, the potential demoralising effects of test results are reduced.
- The criteria by which they are being assessed should be made clear to the learners. This process should begin before the test, when the assessment criteria can be explained. Providing this has been done, the feedback stage is then an opportunity to revisit the criteria and highlight the gap between what the child can currently do and what they are aiming for.
- Feedback should challenge the learner and require some form of (achievable) action. In other words, feedback should not be a final judgement from the teacher that requires no response. If there is no push to reflect on, or respond to, the feedback in some way, it is likely that an opportunity for improvement is being missed. By asking the children to rewrite something based on feedback, for example, we try to ensure that they see progress in their learning.

The third general recommendation is that self-assessment be made part of the teaching programme. This will help children to develop the habit of monitoring their own progress, which in turn can set them on the path to becoming active learners, a quality known to benefit learning in general.

Below is an example of post-test material produced by the Norwegian Ministry of Education for 11–12-year-olds (Hasselgren 1999). Pupils complete the form after doing an assessment task on reading.

EPISODE 1

Try to answer these questions. Put crosses.

Did you ...	yes	mostly	so-so	not really	no
understand what to do?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
understand the texts?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
have enough time?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
do the tasks well?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
like the tasks?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
manage to guess what new words meant?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Were any texts difficult to understand?

☐ no

☐ yes (write the numbers)

What have you learnt?

Our final recommendation is that every effort be made to create the conditions that allow the children to perform at their best. This means, we think, that they should be tested by sympathetic teachers whom they know and in surroundings in which they are familiar. It is particularly important with children to make sure at the outset that they understand what they have to do. With this in mind, it is preferable that a test is in a similar format to those the children have met in class. Or, if this is not possible, there should be practice items at the start of the test or at least one or two examples which model what the child is expected to do. It is also important to include easy tasks at the beginning of a test in order to give them the confidence to tackle the more difficult ones.

Our recommendations and their intended outcomes may seem somewhat idealistic, but before rejecting them one has to consider the alternative; by default, this is to instil negative attitudes towards tests, and, through them, to language learning.

Specific test features

Although we want children to take tests in a relaxed setting, this does not mean that we should relax our own standards for test development. We still need to make sure that our tests are valid and reliable¹.

And the need to seek positive backwash is more important than ever. It would not be appropriate here to recapitulate the advice given earlier on how to make tests valid and reliable, and achieve beneficial backwash. It is worth saying, however, that crucial elements are the writing of full specifications and the choice of appropriate test techniques.

Before considering particular test techniques, let us ask what it is about young learners that might require their test to have specific features.

1. Young learners have a relatively short attention span. For this reason, tests should not be long. Individual tasks should be brief and varied. If necessary, what would for other learners have been a single test can be broken down into two or more tests.
2. Children enjoy stories and play. If we want them to become engaged in tests, the tasks should reflect this. Games can include versions of the kind of word games to be found in comics and puzzle books. Furthermore, it is our experience that children react well to 'silly' scenarios which appeal to their sense of fun. Stories and activities with elements of silliness are to be encouraged.
3. Children respond well to attractive typography, colour, pictures and videos². Tests should include these features where possible. Pictures can serve as options in multiple choice tasks; along with videos, they can be stimuli in speaking and writing tasks. They can also be used as visual support, either to set the scene for an activity or as further explanation of task instructions. Pictures may be included even when they are not necessary to complete a task, in order to make the test less forbidding. It goes without saying that the content of all pictures used should be unambiguous for all the children who may take the test. This might involve testers in checking that children with different cultural backgrounds are familiar with the conventions (such as the different kind of bubbles for speech and for thought) that are used in the test pictures.
4. Most children these days are 'digital natives', and therefore tasks which use digital technology will appear entirely natural to them. The use of a tablet, for instance, allows for the creation of engaging, interactive

¹ Attractive as they might seem for young children, *True/False* and *Yes/No* items, for example, are no more valid or reliable for them than they are for adults.

² Unfortunately, it was not possible to include colour in this book.

tasks, involving features such as ‘click and drag’, with which the child is already familiar from playing computer games. Readily available software facilitates the construction of such tasks.

- 5. First language and cognitive skills are still developing. Tasks should be ones that the children being tested can be expected to handle comfortably in their first language. An analysis by Hasselgreen (reported in Hasselgreen and Caudwell 2016) identified the criteria for each level of the *Common European Framework of Reference (CEFR)* which it would be unrealistic for children of various ages to meet. For example, a successful test-taker at Level B1 should be able to “express thoughts on abstract and cultural topics” and “explain main points in an idea”, both of which require cognitive abilities not commonly found in children under the age of eight or nine.

TABLE 3: CORRESPONDENCE BETWEEN AGE GROUPS AND <i>CEFR</i> LEVELS POTENTIALLY ATTAINABLE	
Age group	Limits of <i>CEFR</i> levels potentially attainable
Young children (roughly between 5/6 years and 8/9 years)	A2 Reading and writing levels will depend on the emergence of literacy
Older children (roughly between 8/9 years and 12/13 years)	B1
Teenagers (roughly between 13 and 17 years)	B2
Exceptional older teenagers	C1

The implications for anyone wanting to align their tests with *CEFR* are clear. More generally, Hasselgreen’s work serves as a reminder to limit test tasks to those which children could be expected to handle in their own language.

Recommended techniques³

In what follows, we have concentrated on techniques that seem particularly suited to young learners. This does not mean that techniques presented in previous chapters will never be appropriate. The older the children are, the more likely they are to respond well to techniques used with adults. Whatever techniques are used with young learners, it is essential that the children have plenty of opportunities to practise with them before they meet them in tests. Ideally the techniques should be used in learning exercises as well as in testing. Many of the examples that follow

³. Children aged eight are quite different from children aged sixteen, and so not all of the techniques given here will be equally appropriate for young learners throughout this age range.

are in a digital format. However, it is important to note that all of these digital examples may be replaced by paper-and-pencil equivalents.

Techniques to test listening

Placing or identifying people, objects and actions⁴

1. In the first example, children see a room with various pieces of furniture on the right of the screen. On the left of the screen are four objects. They hear instructions and attempt to follow them.



The children hear:

"Look at the pictures. Now click and drag."

"Put the teddy bear on the table." (They hear this twice)

Tasks such as this, which require students to 'click and drag', can easily be recreated in a non-digital context by presenting children with a drawing and asking them to 'draw lines' from the object to the correct position in the picture.

2. A second example involves seven images of children taking part in a variety of activities, and the names of seven children.

⁴ These techniques may be seen as varieties of multiple choice, most obviously in the third example. This is unproblematic, provided that the warnings given in the chapter on test techniques about the use of multiple choice are observed.

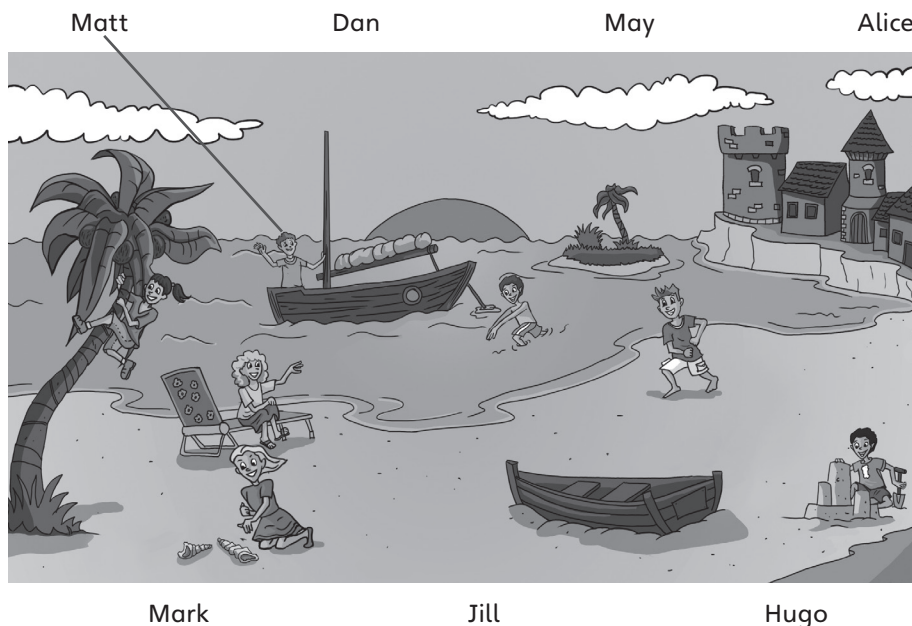
Test 2

Listening

Part 1

– 5 questions –

Listen and draw lines. There is one example.



The children hear:

Look at the picture. Listen and look. There is one example.

Girl: Here's a picture of some kids from school and me, Grandpa.
We're on the beach.

Man: Oh, yes. It's a great picture. Who's that? The boy in the boat?

Girl: That's Matt. But it's not Matt's boat. It's his brother's.

Man: I see.

Can you see the line? This is an example. Now you listen and draw lines.

Man: And who's that girl? The one in the coconut tree?

Girl: That's May. She loves coconuts!

Man: Is May a good friend?

Girl: Yes! She's in my class. I really like her.

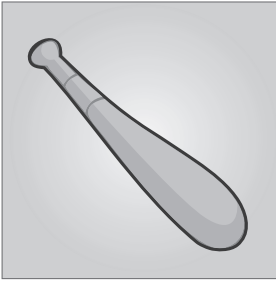
Man: And what's that girl's name?

Girl: That's Jill.

And so on

3. In this third example, the children see three pictures and a written question. They listen to a short conversation and tick the correct box.

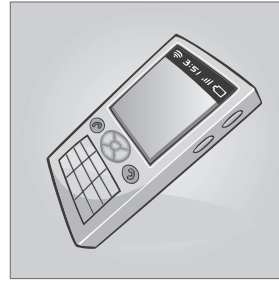
4 What did Nick get for his birthday?



A ☐



B ☐



C ☐

The children hear:

What did Nick get for his birthday?

A: Did you have a good birthday, Nick?

B: Yes! I had some great presents too!

A: And what did your parents give you? A new phone?

B: No, I've got one of those. I wanted a guitar but they gave me a baseball bat.

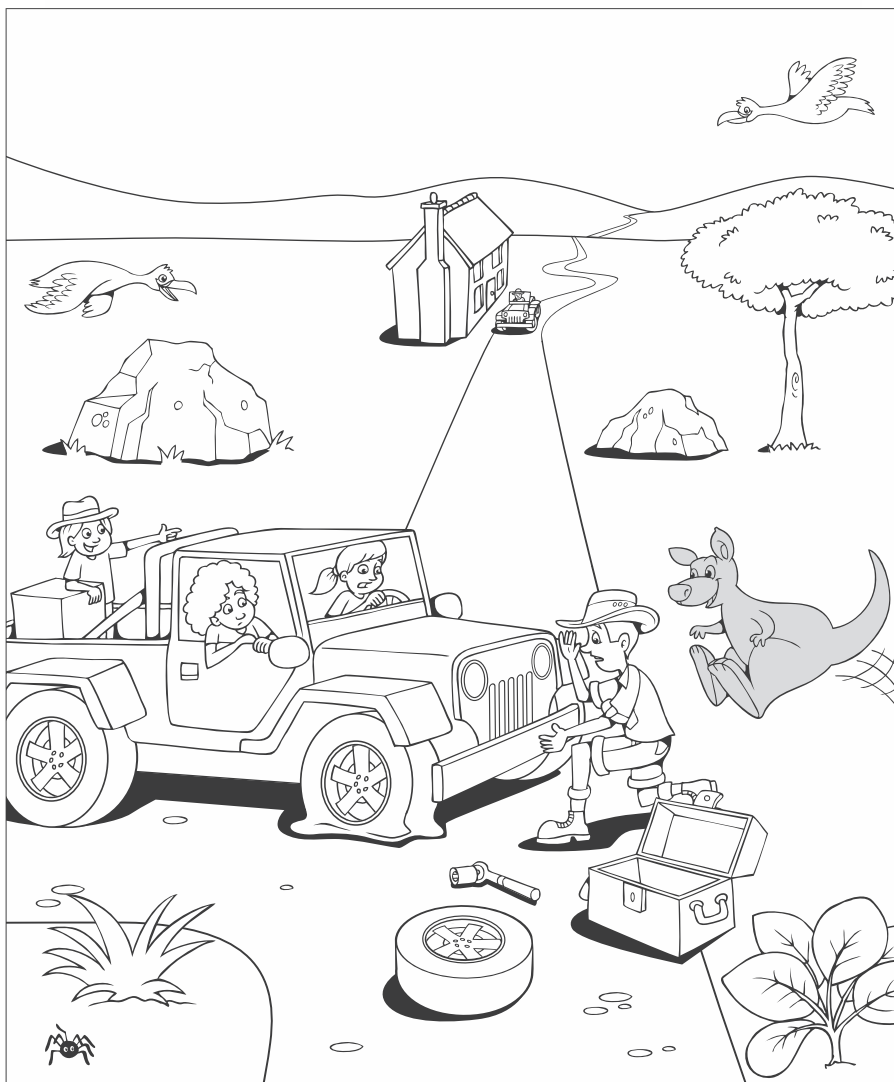
A: That's good!

B: Yes. I needed a new one⁵.

⁵ This item has fewer options than the previous two and is therefore likely to contribute less to test reliability, other things being equal.

Identifying and colouring objects in a line drawing

This type of task can either be digital or make use of paper and coloured pencils. The example is from a *Cambridge Young Learners English Test*.



The children hear a conversation between an adult and a child who are looking at the same picture. After listening to an example, the children hear:

A: Can I colour one of the plants too?

B: Which one? The one with the round leaves?

A: Yes. I like the one with the round leaves the most.

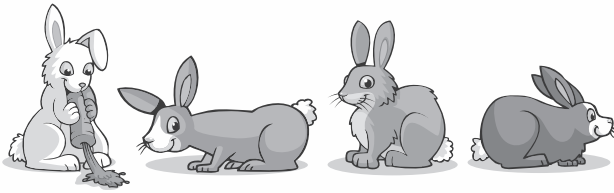
B: Alright. Colour it blue.

A: OK. I like that colour.

And so on.

Information transfer

This may often involve some simple reading and writing. For example, there may be a chart:



Mr Mat's rabbit

Likes drinking: carrot juice

Colour of rabbit:

Bought where: pet shop next to.....

Name of rabbit:

Lives in: Mr Mat's.....

Likes eating: Mr Mat's.....

The children hear a conversation between a child and a teacher. The teacher is telling the child about his pet rabbit. All the information the child needs in order to complete the chart is included in the conversation. With tasks like this, the talk or interview should include sufficient redundancy and include pauses during which answers can be put in the chart.

A: Mr Mat? I want to buy a rabbit for a pet.

B: That's a good idea. I've got a rabbit.

A: Have you? What does your rabbit like to drink?

B: It likes drinking carrot juice.

A: Carrot juice?

B: Yes.

(Pause)

A: What colour's your rabbit Mr Mat?

B: There are lots of different colours of rabbits but mine's grey.

And so on.

Techniques to test reading

Multiple choice

The use of images in multiple choice items in tests of reading means that the children do not have to process two texts in order to demonstrate understanding of one of them.

1. The first example is taken from a Norwegian test for 10-year-olds at *CEFR A2* level.

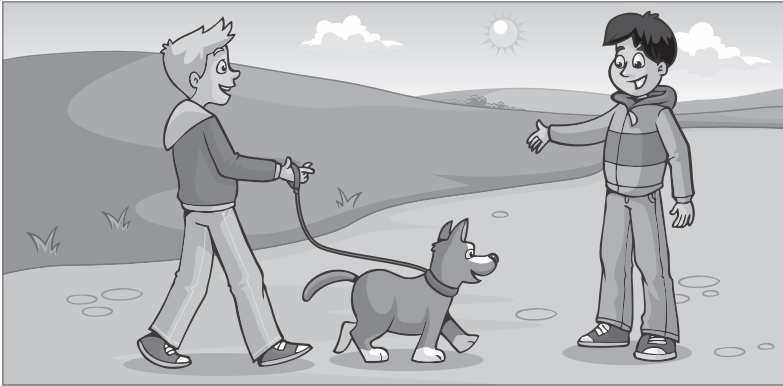
Children read the text and click on the relevant picture.

An English map maker is believed to have been the first to start selling this item, in the mid-1700s. Generally the pieces are made of cardboard now, but at first they were made of wood. It contains many small pieces that must be put together to make a complete picture.



2. Multiple choice can also be used in the context of a transcribed conversation, interview or discussion. The children have to choose the most appropriate response to something that is said. The following item is from the *Cambridge English Young Learners* test. The image serves simply to provide context for what is said.

Read the text and choose the best answer.



Example

Paul: Fred, whose dog is that?

- Fred:**
- A There it is.
 - ☒ B He's mine.
 - C That's new.

3. For older children, the reading comprehension items may involve longer texts, as in the following example.

Read the text. Click on the correct answer.

Alice loves to bake cakes. Alice's brothers, Sam and Oliver, have hobbies that are very different from hers. They love to play video games and watch movies. But their favourite hobby is playing pranks on their sister. This drives her crazy!

One day Alice was baking a birthday cake for their father. After tasting the batter, she realized that her brothers had switched the salt and the sugar! Alice decided that she would find a great way to pay them back.

Later that evening, when the boys were busy playing video games in the basement, she went into their bedroom. She set the alarm clock to go off at 5 am. But Alice decided this was not enough. Next she snuck into their bathroom and poured blue food colouring into their shampoo. She knew that, having woken up so early, they wouldn't notice that the shampoo was blue.

Quite a sight met Alice in the kitchen the next morning. Two tired boys with blue hair sat at the table having their breakfast. Alice and her parents couldn't stop laughing.

Why are Sam and Oliver tired?

- ☐ They had played too many video games.
- ☐ They had stayed up late watching movies.
- ☐ They had been woken up very early.
- ☐ They had washed their hair too many times.

Definitions

Simple definitions can be made the basis of multiple choice reading test items. To reduce the chances of correct responses being made by guessing, a single item may include several definitions. For example, there may be ten definitions and a set of fifteen words (which include ten to which the definitions apply). The children have to identify the correct word and copy it alongside its definition.

The definitions need not be formal. For instance, *pet* may be defined as 'an animal you keep at home'. Provided that the presentation of such items is attractive (the words may be different colours, for example, and dotted around the page or screen), such items need not be as grim as they may first sound.

Children draw a line from an image to the appropriate text

In the following example, there are six images but eight sentences, thereby reducing the possible effects of guessing.

4 Task four A day at the circus.

Zoe, Arlo and Elias go to the circus. Their father takes them there. Draw a line from each picture to the correct sentence.

Be careful. There are two extra sentences.

The first one is an example.



They are so funny.

I don't like elephants.

Have a sweet.

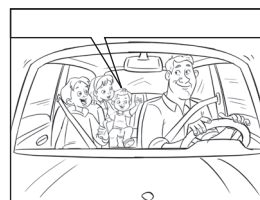
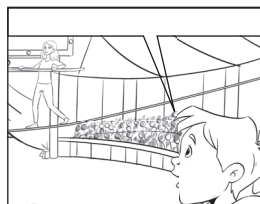
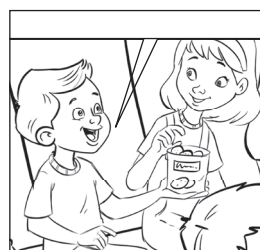
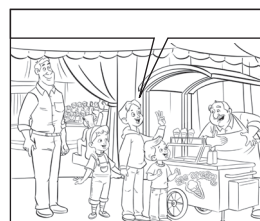
→ I'll be here when you come out.

Oh no, it's raining again!

Can we have three, please?

Don't fall.

Thanks, Dad, we had a good time.



Techniques to test writing

Anagram with picture

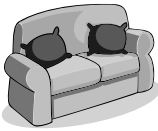
To test vocabulary and spelling, children can be presented with a 'puzzle'. There is a series of pictures and opposite each picture is an anagram of the word the picture represents, as in the following example from *Cambridge English Young Learners* test.

Part 3

– 5 questions –

Look at the pictures. Look at the letters. Write the words.

Example



s o f a



Questions

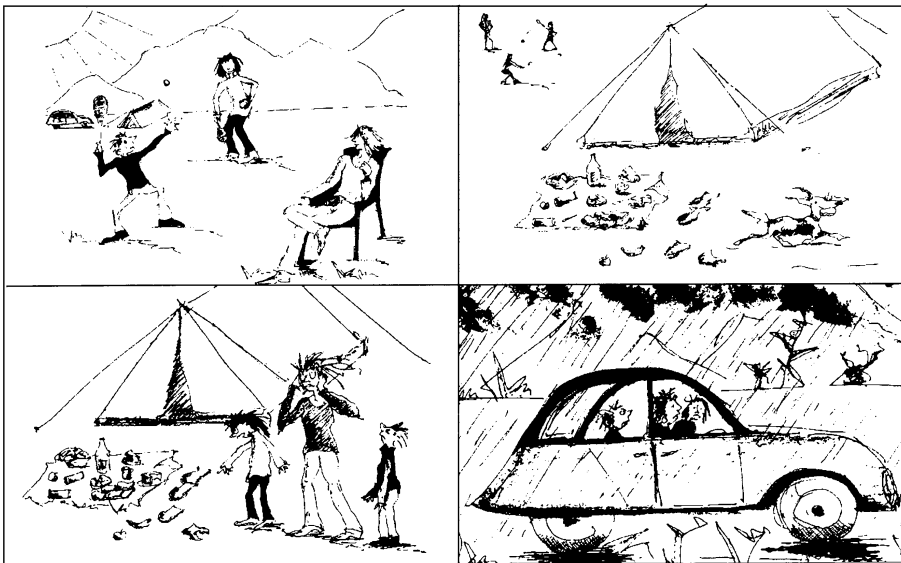
1





Cartoon story

A series of cartoons tells a simple story. The task can be paper-and-pencil, or it can be made digital.



The instructions are:

Look at the pictures. See what happens. The girl in the picture is called Sally. Sally writes a letter to her friend David. She tells him what happened.

Here is her letter. Write what she says to David.

Dear David

*Best wishes
Sally*

Gap filling with pictures

A passage (perhaps a story) is presented in which there are blanks where words are missing. Above each blank there is a pictorial representation of the missing word. Provided that the text is simple and undemanding, the need to read is unlikely to affect performance in writing.

I live in a small



by the sea. Every day I go for

a swim. One day, when I came back after a swim I saw a big



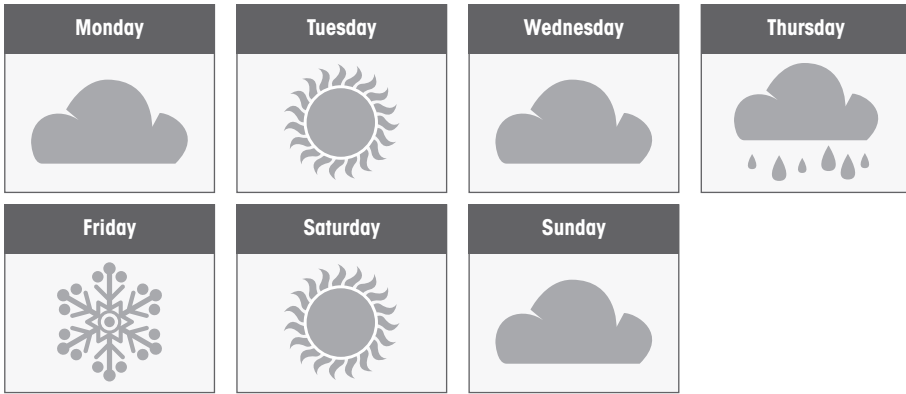
. In its mouth was a big



Information transfer

Here is a simple example.

A chart shows the week's weather (7 days), including the names of the days, and using symbols for rain, sun, cloud and snow. Below the chart are the following sentences which the child has to complete.



On Sunday, there was _____

On Tuesday, there was _____

On Friday, there was _____

(Based on Hasselgreen and Caudwell 2016)

A series of related tasks based on a website⁶

In a website setting, the children are required (for example) to:

1. complete a form with their basic personal details (name, date of birth, etc.);

Teens Writing Tasks Sample Item

NB: All items in *italics* change with each new item to follow the theme.

Task 1

Make friends from around the world with *Global friends*. It's simple to join and fun to use. Fill in the form.

1. Your name	
2. Your date of birth	
3. Your country	
4. Your first language	
5. Your favourite <i>places</i> . List 3	

Now click on the button to go to step two. [button]

⁶ It is not intended that these tasks are actually carried out online. Where it is not possible to construct (what looks like) a website, a paper-and-pencil equivalent is possible, though this is likely to be less engaging of real interest.

2. write a couple of sentences about their friends and things that they like;

Task 2

Tell the other members something about yourself. Fill in the form. Write in sentences. Use 20-30 words.

1. Personal information	What do you normally do with <i>your friends</i> ?
2. Preferences	Which do you prefer, <i>cars</i> or <i>trains</i> ? Why?:
3. Opinions	What do you think about <i>homework</i> ?

Click on the button to register [button]

3. post an online comment on a topic introduced in the task;

Task 3

Welcome to *Global friends*. Use our forum to meet other Teens from around the world.

Miguel from Spain has posted this photo on the forum. Add a comment and then reply to two comments from other members. Use 20-30 words for each comment.



Miguel (Spain): *I took this photo at the annual La Tomatina festival in Valencia, Spain. It's really crazy but lots of fun. Would you like to take part in a tomato fight? Why/Why not?*

You:

[post]

Chie (Japan): *I don't think I'd like it. My mum made me eat tomatoes when I was a little kid and now I can't stand them ;-)* Has anybody else got a food hate like me?

You:

[post]

Sandra (Colombia): *I'd love to have a go! I love doing crazy things. Last summer I went surfing. What kind of extreme sports can you do in your country?*

You:

4. provide a more extended piece of writing on a similar topic (framed as a competition entry).

Task 4

Every month we run a competition on our website. Why not enter? You might win one of our fabulous prizes! The theme this month is 'Global issues'. Write your argument in response to this statement: *'There is no need to recycle or use alternative sources of energy as it will make no difference to global warming'*. Remember to include an introduction and a conclusion.

Write your competition entry in 220-250 words here.

(Based on Hasselgreen and Caudwell 2016)

The website setting allows for the construction of realistic tasks, all related to each other. Note that the tasks can easily be presented in increasing order of difficulty. The above example is intended for teenagers, but simpler tasks could be designed for younger learners who are already used to accessing websites.

Techniques for testing speaking

The same general advice for testing speaking given in Chapter 10 applies equally to the testing of young learners. What is worth emphasising, perhaps, is the need for a long enough warm-up period for the children to become relaxed. In the case of the youngest children, it may be helpful to introduce toys and dolls from the outset.

- Asking straightforward questions about the child and their family.
- Giving the child a card with a scene on it (a 'scene card'), and then asking them to point out people, say what colour something is, what someone is doing, etc.
- Giving the child two pictures that are very similar but which differ in obvious ways (for example, one picture might contain a house with three windows and a red door, with a man in the garden; while the other might have a house with four windows, a green door and a woman in the garden). The child is asked to say what the differences are.
- The child is given a short series of pictures that tell a story. The tester begins the story and asks the child to complete it.
- Sets of pictures are presented. In each set there is one picture which does not 'belong'. There may, for example, be three pictures of articles

of clothing and one of a bed. The child is asked to identify the odd one out and explain why it is different from the others.

- Cards can of course be replaced by tablets (or some other electronic means of display). Tablets are particularly useful in presenting videos (rather than static images) to which the children have to respond.

Where we want to see how well children can interact with their peers, useful techniques are:

- If the two children belong to the same class, each can say a specified number of things about a third classmate, at the end of which the other child has to guess who is being described.
- There are four different picture postcards. Each child is given three of them, such that they have two cards in common and one which is different. By asking and answering questions in turn, they have to discover which pictures they have in common. All the pictures should have some common features, or the task may end too quickly without much language being used.
- There are two pictures (A and B) which are different but which contain a number of objects that are identical. One child is given picture A, the other picture B. The first child has to describe an object in their picture and the other has to say whether it is to be found in their picture. The second child then describes something in their picture, and the other responds. This continues until they have found a specified number of objects which are in both pictures.
- The children can each be given a card with information on it. In both cases the information is incomplete. The task is for them to ask questions of each other so that they end up with all the information. Examples would be diaries with missing appointments, or timetables with missing classes. A variant of this is the Question and Answer Board Game, which at the time of writing forms part of Pearson's *PTE Young Learners*.



READER ACTIVITIES


1. Find Rixon's survey online, using the following search terms 'Rixon survey British council young learners'. Look at the levels young learners are expected to reach at the end of primary school in the following countries: Colombia, Czech Republic and Denmark.


How well do these expected levels fit with Hasselgreen's analysis?


Ask the same question about your own country or a country with which you are familiar.


2. Look at the following activities taken from *Power Up* (Nixon and Tomlinson 2018). These were not originally devised as test tasks. What changes, if any, would you make to them in order to create test tasks that will be reliable and valid?


1 Find and circle the words. Then write.



 head







































w	n	o	s	e	y	e	m
y	m	o	u	t	h	a	d
k	l	e	g	a	a	r	m
f	o	o	t	i	n	l	h
e	r	f	y	l	d	o	e
e	s	a	h	a	i	r	a
t	q	c	b	o	d	y	d
l	y	e	t	c	e	i	h


1 Listen and join. Then write.



 Hugo



 Sam



 Pat



 May



 Tony



 Alex



 a


 b


 c


 d


 e


 f

- 1 Hugo can play tennis.

2 _____

3 _____

4 _____

5 _____

6 _____



FURTHER READING

Cameron (2001) is a book on teaching language to young learners, which has a chapter on assessment. Ioannou-Georgiou (2003) and McKay (2006) both offer practical advice on the assessment of young learners.

Rea-Dickens and Rixon (1997) discuss the assessment of young learners of English as a foreign language. Carpenter et al. (1995) describe an oral interview procedure for assessing Japanese as a second language. Hasselgreen and Caudwell (2016) is a book devoted to the subject of the assessment of young learners. Hasselgreen has a chapter on the assessment of young learners in Coombe et al. (Eds 2012a).

Language Testing 17, 2 (2000) is a special issue on assessing young language learners. Contributions include a general introduction to the area by Rea-Dickens; an account of how foreign language attainment is assessed at the end of primary education in the Netherlands by Edelenbos and Vinjé; a discussion of teacher assessment in relation to psychometric theory by Teasdale and Leung; a description of the Norwegian materials project (referred to in the chapter) by Hasselgren.

Sample papers for Cambridge and Pearson tests for young learners are available free online.

16

Beyond testing: other means of assessment

Testing is the focus of this book. The clue is in the title! In Chapter 1 we argued for the necessity of language testing. At the same time, however, we have said that other forms of assessment may often be more appropriate, particularly in formative assessment.

Classroom assessment is a topic worthy of a book in itself. In this chapter, we can only offer an outline of what is possible. The reader interested in looking more deeply into the topic will find many suggestions in the Further reading section.

We will begin by discussing the role played by less formal assessment; then we will identify principles for its implementation; and finally, we will describe a number of methods by which it can be carried out.

We see less formal assessment as making the following contributions:

1. It provides the teacher with continuous diagnostic information as to what has been learned and what has still to be learned. This information can then be used to provide remedial work for individuals, for groups of students, or for the whole class. Where it indicates that most students have failed to learn something that has been taught, this information can prompt the teacher to consider possibly more effective ways to teach it.
2. It can provide feedback to students as to how well they are doing. It lets them see what progress they have made.
3. If the feedback is expressed in relation to short- and long-term course objectives, it will help make students aware of those objectives.
4. It can encourage learner autonomy, allowing students to take control of, and accept responsibility for, their own learning. Their active involvement in the entire assessment and remediation process can lead them to a better understanding of the language learning process.
5. Awareness of the progress they are making towards these objectives may promote students' intrinsic motivation (as opposed to the extrinsic motivation of examination results and their academic and social consequences).

If, on the basis of feedback, students are involved in making a plan for remedial work (individually or in groups), this may further promote motivation.

6. Finally, information obtained by informal methods of assessment may be used to supplement test results. This can be especially useful in the case of students with borderline test scores. More generally, where there is a large discrepancy between a candidate's test score and other measures of their language ability, further investigation is called for. Ideally, there should be congruence, though normally some inconsistencies between test scores and other assessments are to be expected. If there are too many significant differences, there is a need to reflect on the validity of both final tests and continuous assessment.



GENERAL PRINCIPLES OF CLASSROOM ASSESSMENT

- The nature and frequency of assessment will depend in part on class size.
- It can be carried out in class or online, with individual students or with groups of students.
- Whatever method is used, assessment should be informal, and integrated as far as possible into regular classroom activities. It should in no way be threatening.
- Assessment sessions should be carefully prepared. Otherwise, they can become disorganised and fragmented¹.
- Feedback should be given immediately and with an emphasis on positive features of performance.
- Careful records should be kept.

Methods of assessment

Observation of performance

This is the principal and most obvious method of classroom assessment. It reflects what normally happens during teaching, except that it is consciously designed to obtain information. Focusing on particular aspects of language related to instructional objectives, the teacher attempts to elicit behaviour that will show whether or not those objectives have been achieved. For instance, to discover if students have mastered the past tense of certain irregular verbs, the teacher may ask a question concerning an event in the past, and then require them to ask one themselves of another student. The teacher makes an evaluation of the performance of each student and records this. Depending on outcomes, the teacher may conclude that only a small number of students have problems (and provide them with help) or that most of the students do (in which case further more general instruction or practice may be thought necessary). This is just one example but we trust that the reader will easily imagine others.

¹. Of course, this does not preclude the possibility of teachers deciding from moment to moment to check whether something has been learned (but without necessarily recording the outcome for future use).

The teacher should keep a careful record of such observations, and make them available for comparison with formal test results. A simple pro forma for recording speaking task assessments might be as follows.

Features	Unsatisfactory	Satisfactory	Outstanding
Individual sounds			
Stress and intonation			
Fluency			

Conferences

These are meetings with one or two students in which the teacher gives feedback on observed performance (or, probably better, teacher and students discuss feedback which has already been provided, and which the students have had time to read and assimilate). By discussing the feedback and reactions to it, the teacher can learn more about the students. Students can explain why they wrote or said what they did. Conferences can also be used to elicit further performance. The student may, for example, be asked to read a text aloud and then answer questions on it. In order to encourage active participation in the learning process, students should be asked to come to conferences with questions and comments.

Presentations

Students are asked to prepare a short talk on some relevant subject and present it to the class. This gives the teacher the opportunity to assess various aspects of speaking ability using a rating scale.

Journals

Writing ability can be assessed through the use of journals in which students reflect on their language learning experience, and indeed on any aspect of their lives. When the teacher and a student both contribute to one journal, responding to each other's entries, it is referred to as a dialogue journal. Dialogue journals can give students the sense of using the new language as a means of authentic communication. Journals can be written on paper or online. They may be between a teacher and a single student, or between a teacher and a group of students.

Projects

Projects require students to create something individually or in collaboration with other class members. They typically involve the integration of all language skills, and call upon skills beyond the purely linguistic (such as critical thinking and problem solving). Achievement may be assessed while the project is in progress or on its completion. They can be the basis for presentations (above).

Online and computer-based programs

Various programs provide students with the opportunity to have language ability assessed. One example, *DIALANG*, was mentioned in an earlier chapter. The Cambridge English's online program *Write & Improve* gives immediate feedback on any piece of writing a student presents, and allows re-presentation after the student has attempted to improve it. Pearson's *English Benchmark* allows teachers to test young learners in speaking, listening, reading, and writing on a tablet. *Socrative* is a popular interactive program which allows teachers to input their own quiz items.

Self assessment

Students are asked to assess their own performance or their current level of ability on simple rating scales of some form or other. Individual students then meet with their teacher to compare their own assessment with the teacher's, discussing points of discrepancy between the two assessments. The teacher's explanation of the discrepancies should help students develop a true picture of their ability.

Rating scales relating to learning objectives (short or longer term) may be developed by teachers. Alternatively, publicly available scales may be used, such as the Council of Europe 'Can do' scales (See the Online resources section).

Peer assessment

Students comment on the work (written or spoken) of fellow students. They typically exchange papers and work in pairs or small groups, commenting and making suggestions for improvement. The benefits are the opportunity to speak the language for a real purpose and with someone other than the teacher as interlocutor. It should also help develop learner autonomy. For peer assessment to work well, it is important that the purpose of the task is well explained and that the students are trained in asking relevant questions and in giving advice in a collaborative fashion. They may be given a pro forma to follow while reading another student's work.

Pop quiz

The teacher asks a small number of questions to the whole class. The students write their answers individually. They exchange their answers with another student, who marks them as the teacher gives the correct answers. The students return their papers to each other and compare their responses, trying to understand any errors they may have made and, where necessary, ask the teacher for explanations.

Portfolios

A portfolio is a folder or binder containing samples of a student's work written over a period of time. It is kept in the classroom and the student

has access to it at all times. It is the student's responsibility to maintain the portfolio, inserting what they think is their best work. It should remind the student of the progress they are making. Portfolios may also be constructed and maintained online (these are sometimes referred to as 'e-portfolios'). Portfolios, together with records of other classroom assessment information, are available for comparison with test scores.

A word of warning

Particularly where important decisions are to be made, we strongly advise against the use of classroom assessment to the exclusion of testing. Where there are no tests, or where test scores are given only a minor role, there is a danger of a too cosy teacher-student relationship developing. This can lead to a misleading picture of student attainment, which benefits no one.



READER ACTIVITIES

1. How would you convince a hard-nosed language tester that the results of other means of assessment should be taken into account when making important decisions about students?
2. Imagine that you want to assess your students' ability with respect to a particular language structure. Choose a structure and then say how you would elicit performance in class. Design a simple chart to record your assessments.
3. Look at the various rating scales presented in earlier chapters. Which of them do you think could be the basis for creating a self-assessment instrument? What modifications would you need to make?
4. Take the various methods of assessment that we have advocated and place them in order according to the usefulness of the information they may provide. Compare your order with that of a fellow teacher who has done the same thing, and discuss.
5. Find *Socrative* online. How useful do you think it might be for assessment purposes?



FURTHER READING

General

For advice on choosing the right type of assessment, see J. D. Brown (2012). *Genesee and Upshur* is a book on classroom-based assessment (1996). Cheng and Fox (2017) is a more recent book on the same subject. *Language Testing* 18, 4 (2001) is a special issue on alternative assessment. *Language Testing* 21, 3 (2004) is devoted to the topic of diversity in teacher assessment. Tsagari (2016), available free online, is a collection of papers on classroom-based assessment. Al Mahrooqi and Denman (2018) discuss non-testing assessment in general, as do Coombe, et al. (2012b).

Methods of assessment

For the role of observation in assessment, see Quirke (2018a), Shehadeh (2012) on task-based assessment. For the use of projects in assessment, see Tuzlukova (2018). For the use of journals, see Quirke (2018b). Sadhwani and Sheetz (2018) discuss presentations. For peer assessment, see Cheng and Warren (2005), Saito (2008), Matsuno (2009), Anderson (2012), Suzuki (2015), Sun and Doman (2018). For self assessment, see Luoma and Tarnanen (2003), Matsuno (2009), Butler and Lee (2010), Anderson (2012), Babaii et al. (2016), Midraj (2018). For portfolios, see Curtis (2018).

Online resources

There are many online self-assessment tests. These include: Cambridge English's *Write & Improve* and *Test Your English*, *DIALANG*, and a free British Council test of English.

To find others, search for 'Online language assessment tools'.

17

New technology and language testing

New technology is having a great impact on most aspects of our lives. This is increasingly true of language testing. Throughout the book, we have drawn attention to some innovations, but we thought it worthwhile adding a chapter where we would reflect on the increasing influence of new technology and attempt to evaluate its current and possible future effects in terms of validity, reliability, backwash and practicality. For the moment, some innovations apply only to major testing organisations but this may change. Teachers should be aware of what is possible.

We will begin by identifying what we regard as the most significant developments in technology that have affected language testing, or are likely to affect it in the future.

Significant developments in technology

Probably the most fundamental development has been the increase in the speed of computer processors. In less than the time the second edition of this book was in print, the speed of the fastest processors went from 9,726 MIPS (millions of instructions per second) to 304,519 MIPS. While it has been suggested that the growth in speed may diminish in future years, there is no reason to think that speeds will not increase further. At the same time, there has been a parallel increase in the speed and power of graphics cards. Massive data sets can be stored, managed and processed on networks of remote servers hosted on the internet (cloud computing).

A second important development has been miniaturisation, which means that smartphones and tablets have many times the computing power of desktop computers of not so long ago, while nanotechnology offers the potential for new and even faster kinds of computer.

A third development has been in telecommunications. The speed with which information can be transferred between devices throughout the world has increased dramatically. At the time we are writing this, 5G (fifth generation) wireless cellular technology is being introduced into the United Kingdom. This will have a theoretical download speed of 10,000 Mbps (Megabytes per second) and will allow more complex apps to be used and processes to be carried out with less hardware. The number of people with access to the internet has grown enormously since the publication of the second edition of this book (from around 10% of the global population to over 50% at the time of writing this).

A group of related developments are in AI (artificial intelligence), machine learning and natural language processing (NLP). AI is the branch of computer science attempting to build machines capable of intelligent behaviour, while machine learning has been defined as the science of getting computers to act without being explicitly programmed. NLP is concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyse large amounts of natural language data. We see the results of work in these fields in, for example, virtual assistants, expert systems which diagnose illnesses, and in automated computer translation services.

Application to language testing

Individualisation of testing

In computer adaptive testing, faster processors mean that calculations can be made at extremely high speed, allowing the reliability of an individual's score in real time between items. In this way, not only can the difficulty of the next item be determined, but the test can be ended as soon as a high enough reliability coefficient for the individual's performance is reached. This means that no one takes more items than is necessary, resulting in a test which is both reliable and more practical than one where everyone has to take all items.

Machine learning is the basis for the automated scoring of both writing and speaking. Using AI, natural language processing and machine learning software, the computer learns simply by being presented with thousands of examples of performance and the 'correct' score for each of them (this score being arrived at by combining the scores of hundreds of experienced human scorers), without the need for the computer to be instructed as to how it should arrive at a score. Automated scoring is clearly reliable, and it is also practical, since human scorers do not have to be on hand to score written and spoken performances whenever an individual takes a test. Test results can be reported instantly. Immediate feedback, particularly where this provides candidates with diagnostic information, is likely to promote positive backwash.

There are, however, issues about validity. The performances which are scored are based on the candidates' interaction with a computer, and in speaking tests this cannot at present include interaction with peers or, indeed, genuine oral interaction of any kind (resulting in a lack of content validity). The computer's inability to understand meaning (in writing as well as in speech) has also been cited as a validity problem, though proponents of automated scoring can point to the fact that the human raters whose scores were used in the learning process did take meaning into account, and that automated scores have agreed with those of raters

as highly as have scores between raters. While some may question the appropriateness of using automated scoring in high-stakes tests, in low-stakes tests it would seem to be clearly a force for good.

High-speed processors and graphics cards, miniaturisation, and developments in telecommunications now permit even complex tests to be taken on laptops, tablets or even smartphones. In principle, tests can be taken anywhere and at any time by any individual with a connection to the internet. Oral interviews can be conducted using, for example, Skype, and indeed this has become increasingly common in placement testing. Without the need for special equipment or supervising personnel, the practicality of such testing is obvious. There may be security issues, of course, but see below.

Authenticity and variety of materials and tasks

The internet allows access to a vast array of texts, images and videos¹.

Their use can make an important contribution to authenticity (and so validity) in language testing. In the case of videos, this has become possible only because of higher-speed processors and graphics cards.

Similarly, computer actions such as click and drag (and its touchscreen equivalent) can readily be integrated into language tests. Their inclusion will also add authenticity for a new generation of language learners, digital natives, for whom using a tablet or smartphone is second nature, and the means by which so much of their communication takes place. As new computer practices develop they can be incorporated into language tests.

Integration of teaching and testing online

It seems likely that in the future more teaching will be computer-based. This will provide the opportunity for teaching and testing to be integrated, with new learning being tested and the results of that testing being used to make decisions about what should come next (whether, for example, to move on to something new or to provide practice in what the testing reveals to be imperfectly assimilated). The potential for positive backwash is obvious. Attractive commercial computer packages which integrate teaching and testing are already available. In principle, there seems no reason why teachers given the right training and authoring systems should not themselves produce them for their institutions.

Cloud computing and innovative software allow educational entities to analyse enormous data sets ('big data') with the aim of discovering the factors underlying learning success as measured by assessment outcomes, and on the basis of these to take steps to improve learning by groups and individuals. Many unconnected data sets (such as attitudes to language

¹ Myriads of online tests are also available. The practicality of using these is obvious, but the reliability and validity of these tests will vary. Teachers need to exercise care in choosing tests which are fit for purpose generally and for their teaching situation in particular.

learning, opinions on coursebooks, etc.) can be analysed all at the same time, and relationships established.

Test security

When a test is taken online, the question arises as to who is in fact taking the test. On low-stakes tests this may not be important. On high-stakes tests, however, it is essential to know that the person who will be awarded a certificate is in fact the one who took the test. Various techniques are currently used to ensure that this is the case. These include:

- facial recognition, a technology which verifies a person from an existing digital image by comparing selected facial features.
- palm scanning, in which the hand is held over a sensor. Unlike fingerprint scanning, it involves no physical contact.
- digital signatures. Candidates sign their completed test with a previously established unique digital signature.
- CCTV cameras installed in centres where tests are taken.

With the development of item banks in which the properties of every individual item are known, it is possible to create randomised test formats, which mean that every test-taker can be presented with a different version of the test. This prevents candidates conspiring to create a copy of a test they have taken in order to disseminate it amongst future candidates.

Future prospects

Two areas of technology which we haven't mentioned so far are virtual reality and robotics. Virtual reality is already being used in language teaching but its use in testing is clearly a future possibility, adding authenticity. The day may also come when candidates will be required to interact with robots.

As for the uses of technology we have identified above, we feel quite confident that individualised testing and the use of automated scoring will increase over the coming years. How far these penetrate into teacher-made tests will depend on how accessible and easy to use is the necessary software, and how ready the teaching profession is to embrace the new technology. However, we suspect that the probable increase in the use of computer-based language teaching with assessment as a component will dispose teachers to acquire the skills to write their own integrated teaching-testing computer programs.



READER ACTIVITIES

1. Consider the various developments that we have outlined above. Are you aware of others that may have occurred since we wrote this chapter?
2. Which do you most welcome, and why?
3. Are there any which are unwelcome? Why?
4. Which do you think will have most effect on language testing in the coming years?



FURTHER READING

Our recommendation for further reading in a rapidly changing field is simply to search online, using search terms taken from this chapter. For information on automated scoring, we suggest adding 'Pearson' and 'ETS' (Educational Testing Services), pioneers in the field.

18

Test administration

The best test may give unreliable and invalid results if it is not well administered. This chapter is intended simply to provide readers with an ordered set of points to bear in mind when administering a test¹.

While most of these points will be very obvious, it is surprising how often some of them can be forgotten without a list of this kind to refer to. Tedious as many of the suggested procedures are, they are important for successful testing. Once established, they become part of a routine that all concerned take for granted².

Preparation

The key to successful test administration is careful advance preparation. In particular, attention should be given to the following:

Materials and equipment

1. Organise the printing of test booklets and answer sheets, or uploading of tests, unlock codes, and other computer-associated test content, in plenty of time. Check that there are no errors or any faulty reproduction.
2. If previously used test booklets are to be employed, check that there are no marks (for example, underlining) left by candidates.
3. Number all the test materials consecutively; this permits greater security before, during and after test administration.
4. Check that there are sufficient keys for scorers, and that these are free of error.
5. Check that all equipment (computers, interactive whiteboards, loud-speaker system, etc.) is in good working order in plenty of time for repair or replacement.

¹. Our advice is directed to those who are responsible for administering institutional tests. External test providers will have their own procedures.

². We cannot hope to predict all the ways in which candidates may seek to gain an unfair advantage. Recent incidents in our experience have included writing on the inside of clothing and on the inside of water bottle labels, and the use of Smart watches. We can only suggest that test administrators keep abreast of developments by searching online, using such terms as 'Ways to cheat in an exam'.

Examiners

6. Detailed instructions should be prepared for all examiners. In these, an attempt should be made to cover all eventualities, though the unexpected will always occur. These instructions should be gone through with the examiners at least the day before the test is administered. An indication of possible content can be derived from the Test administration section, below.
7. Examiners should practise the directions that they will have to read out to candidates.
8. Examiners who will have to use equipment (for example, interactive whiteboards) should familiarise themselves with its operation.
9. Examiners who have to read aloud for a listening test should practise, preferably with a model audio-recording (see Chapter 12).
10. Speaking examiners must be thoroughly familiar with the test procedures and rating system to be used (only properly trained speaking examiners should be involved).

Invigilators (or proctors)

11. Detailed instructions should also be prepared for invigilators, and should be the subject of a meeting with them. See the Test administration section, for possible content.

Candidates

12. Every candidate should be given full instructions (where to go, at what time, what to bring, what they should do if they arrive late, etc.).
13. There should be an examination number for each candidate.

Rooms

14. Rooms should be quiet and large enough to accommodate comfortably the intended number of candidates. There should be sufficient space between candidates to prevent copying.
15. For listening tests, the rooms must have satisfactory acoustic qualities or, for computer-based listening tests, all candidates will require individual headphones.
16. The layout of rooms (placing of desks or tables) should be arranged well in advance.
17. Ideally, in each room there should be a clock visible to all candidates.

Administration

18. Candidates should be required to arrive well before the intended starting time for the test.

Test administration

19. On arrival, candidates should be instructed to switch off phones and place them in their bag. All bags are then put together in a corner of the room.
20. Candidates arriving late should not be admitted to the room. If it is feasible and thought appropriate, they may be redirected to another room where latecomers (up to a certain time) can be tested. They should certainly not be allowed to disturb the concentration of those already taking the test.
21. The identity of candidates should be checked.
22. If possible, candidates should be seated in such a way as to prevent friends being in a position to pass information to each other.
23. The examiner should give clear instructions to candidates about what they are required to do. These should include information on how they should attract the attention of an invigilator if this proves necessary, and what candidates who finish before time are to do. They should also warn students of the consequences of any irregular behaviour, including cheating, and emphasise the necessity of maintaining silence throughout the duration of the test. If listening and speaking are being tested in a computer-based test, candidates should be asked to check sound before they start the test, and inform an invigilator of any problem. Invigilators will need to have been told what to do in such cases.
24. Test materials should be distributed to candidates individually by the invigilators in such a way that the position of each test paper and answer sheet is known by its number. A record should be made of these. Candidates should not be allowed to distribute test materials.
25. The examiner should instruct candidates to provide the required details (such as examination number, date) on the answer sheet or test booklet, or, in the case of a computer-based test, to enter their logins/ unlock codes and check their details are correct.
26. If spoken test instructions are to be given in addition to those written on the paper/screen, the examiner should read these, including whatever examples have been agreed upon.

27. It is essential that the examiner time the test precisely, making sure that everyone starts on time and does not continue after time.
28. Once the test is in progress, invigilators should unobtrusively monitor the behaviour of candidates. They will deal with any irregularities in the way laid down in their instructions.
29. During the test, candidates should be allowed to leave the room only one at a time, ideally accompanied by an invigilator.
30. Invigilators should ensure that candidates stop work immediately they are told to do so. Candidates should remain in their places until all the materials have been collected and their numbers checked.

19

The statistical analysis of test data

'There are three kinds of lies: lies, damned lies and statistics,' Benjamin Disraeli supposedly said. It's true that statistics, when misused, may hide the truth. But statistics can also be enlightening.

The purpose of this chapter is to show readers how the analysis of test data can help to evaluate and improve tests. Note the word 'help'. Statistical analysis will provide the tester with useful information that may then be used in making decisions about tests and test results. But it does not take those decisions. This remains the tester's responsibility and depends not only on the information that statistical analysis provides but also on judgement and experience.

The emphasis throughout the chapter will be on the interpretation of statistics, not on calculation. In fact it will be assumed that readers who want to analyse their own tests statistically will have access to computer software that will do all the necessary calculation (see end of chapter for software package suggestions). There is no reason these days to do this calculation by hand or to write one's own programs to do it. For that reason, we have not thought it necessary to show any calculations except the most simple, and these only as part of the explanation of concepts. Where the concepts and calculation are more complex, for all but a small minority of readers the inclusion of calculations would only confuse matters.

There is no pretence of full coverage of the statistical methods and issues related to testing in this chapter; that would take a book in itself. Rather, the basic notions are presented in a form which it is hoped will be recognised as both accessible and useful.

There are essentially two kinds of statistical information on tests. The first relates to the test as a whole (or sometimes to sections of a test); the second relates to the individual items that make up the test. This chapter will deal with each of these in turn, first using a single set of data on a *norm-referenced placement* test, before turning briefly to the analysis of *criterion-referenced* tests. The placement test, which we will refer to as OURTEST, has 100 items and was taken by 186 people.

Analysis of the test

Frequency tables

One begins test analysis with a list of the scores made by each individual taking the test. In the present case this means we have 186 scores. A list of

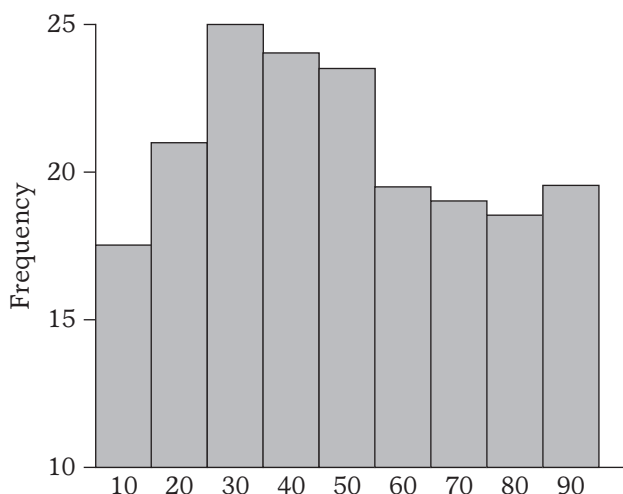
186 scores is not very helpful in understanding how the people performed on the test. A first step in getting to grips with the data is to construct a frequency table. Here is part of the frequency table for the placement test.

Score	Frequency
15	6
16	0
17	2
18	1
19	4
20	6
21	2
22	3
23	2
24	4
25	3
26	2
27	3
28	2
29	2
30	2
31	4
32	2
33	2
34	2
35	2
36	5
37	2

The frequency table tells us that six people scored 15, nobody scored 16, two people scored 17, and so on. Frequency tables are useful when we are considering the effects of different possible cut-off points or pass marks. We can see how many people will pass, fail, or be categorised in some other way (given a particular letter grade, for example).

Histograms

It is still difficult, however, to get a general picture of performance from a frequency table, especially when there are a large number of different scores. In order to get this general view of performance, the frequency distribution can be condensed into what is called a histogram. The histogram for OURTEST appears below.



OURTEST

The diagram should be self-explanatory: the vertical dimension indicates the number of candidates scoring within a particular range of scores; the horizontal dimension shows what these ranges are. It is always wise to create a histogram in order to be made aware immediately of features of the distribution (for example, many people scoring high and low, but relatively few scoring in the middle).

Measures of central tendency: the mean, the mode and the median

Once one has a general picture of performance, a next step is to find what one might think of as a 'typical score'. The most commonly used of the typical scores (also known as *measures of central tendency*) is the *mean*. The mean is simply the average of all the scores made on the test. Add up everyone's score on the test and divide by the number of people taking the test – and you have the mean score on the test.

6 people take a test

Their scores are: 27, 34, 56, 56, 75, 81

The total of their scores is $27 + 34 + 56 + 56 + 75 + 81 = 329$

329 divided by 6 = 54.83 which is the mean score on the test.

We are told that the mean score on OURTEST is 41.56.

The other measures of central tendency are:

- The *mode*, which is the most common score. The mode of the scores in the test above is 56.
- The *median*, which can be found by putting all the individual scores in order of magnitude, and choosing the middle one. In the test above the median is 56, the same as the mode. (As there is an even number of test takers, there are two middle scores. In such cases one takes the two middle scores, adds them together and divides by 2. Here that means adding 56 to 56 and dividing it by 2.)

Normally we can expect the mean, median and mode to be similar to each other. If there are big differences between them, this alerts us to the fact that something unusual has happened, which should be investigated. For this reason, it is a good idea to look at these measures of central tendency at the outset.

Measures of dispersion: the standard deviation and the range

The mean by itself does not always give an adequate summary of a set of scores. This is because very different sets of scores may have the same mean. For example, one group of five students may score as follows on a test: 48, 49, 50, 51, 52. Another group of five students taking the same test may score: 10, 20, 40, 80, 100. Although for each group the mean score is 50, the distribution of scores is quite different. One set of scores is clustered close to the mean; the other set of scores is more spread out. If we want to compare two such sets of scores, stating the mean alone would be misleading.

What we need is an indication of the ways the scores are distributed around the mean. This is what the *standard deviation* gives us. Just as the mean can be seen as a 'typical' score on a test, the standard deviation can be seen as a typical distance from the mean. We do not think it is worthwhile showing how to calculate the standard deviation here; calculation is tedious by hand.

The standard deviation on OURTEST is 23.90.

Another useful measure of dispersion is the *range*. The range is calculated by subtracting the lowest score anyone made on the test from the highest that anyone made. Thus, if the lowest score was 15 and the highest was 86, the range would be $86 - 15 = 71$. The range on OURTEST is 86 ($88 - 2$).

Reliability

We know the meaning and significance of reliability from Chapter 5¹. It was said there that there are a number of ways of calculating the reliability coefficient. Each way is likely to give a slightly different coefficient. For the data we are looking at, we are given four coefficients, which range from 0.94 to 0.98. Without needing to understand the difference between these coefficients, one could quite happily choose the lowest of them, knowing that it is the least likely to be an overestimate. If, of course, one were hoping to sell the test, one might be tempted to choose the highest coefficient!

What all of these estimates given have in common on this occasion is that they are based on people taking the test only once (see Chapter 5 for the

¹ As we also indicated in Chapter 5, for criterion-referenced tests, the equivalent to reliability is 'decision consistency'. We showed how to calculate this.

rationale for this). The tester has divided the test into two halves, which are believed to be equivalent. In the present case, one half is made up of the odd-numbered items, and the other half is made up of the even-numbered items.

Reliability coefficient 1 = 0.94

This coefficient is calculated using Analysis of Variance (or ANOVA). It takes into account the fact that, despite the tester's attempt to create two equivalent half-tests, the actual means and standard deviations of those tests are different. The mean of one half is 19.88, while the mean of the other is 21.68. The standard deviation of one is 12.57, while that of the other is 11.59.

Reliability coefficient 2 = 0.95

This coefficient is also calculated using ANOVA. Because it ignores the difference between the means and deviations of the two half-tests, it is slightly higher.

Reliability coefficient 3 = 0.98

This coefficient is arrived at by first calculating the correlation between scores on the two half-tests (which is 0.96) and then applying what is known as the Spearman-Brown Prophecy Formula. The two half-tests are (obviously!) shorter than the whole test. We know that the longer the test is (if the items are of the same quality), the more reliable it will be. The Spearman-Brown Prophecy Formula estimates the effect on the correlation coefficient of doubling the length of the test.

Reliability coefficient 4 = 0.98

This coefficient is based on the Kuder-Richardson Formula 20. It compares the proportion of correct and incorrect responses on each item. The key thing to remember is that this coefficient is equivalent to the average of all the coefficients that could be calculated using the method that resulted in Reliability coefficient 3.

The reliability of OURTEST is high. If it is thought to be higher than necessary, in order to have a shorter and more practical test (in terms of time to administer and score) one could think of removing items from the test. As OURTEST test was intended for low-stakes placement, a second version of the test was created by removing 40 items out of the original 100. The reliability of the shorter version remained high, at around 0.90. How items should be chosen for removal from a test is explained below.

If the reliability of a test is considered to be too low, one possibility is to add items to it. But if the test already has 100 items and isn't reliable enough, this is hardly a sensible course of action. One needs to look closely at all aspects of the test in its present form, including the way it is administered, and think how it might be made more reliable. Advice in doing this was given in Chapter 5.

The Standard Error of Measurement

We know from Chapter 5 that the Standard Error of Measurement (SEM) allows us to make statements about a person's true score in relation to the score they actually obtained on the test. Other things being equal, the greater the reliability, the smaller the SEM². Taking the lowest estimate of reliability (which is 0.94), the SEM of OURTEST is 2.90.

Knowing that the SEM is 2.90, we can make the following statements:

If someone scores 40 on the test we can be 68 percent certain that their true score is between 37.1 and 42.9 (that is, 40 plus or minus SEM);

And we can be 95 percent certain that their true score is between 34.2 and 45.8 (that is, 40 plus or minus $2 \times \text{SEM}$).

As was said in Chapter 5, the SEM provides information which is helpful when we have to make decisions about individuals on the basis of their performance on a test. It also helps us to decide whether or not our test is sufficiently reliable.

Before moving on to the second section of this chapter, readers might like to look at the following output, and assure themselves that they understand it.

Overall test mean is 41.56 with standard deviation 23.90

Reliability analysis of data in the file OURTEST

There were results from 186 people

Responding to a total of 100 items

First test (part): Mean = 19.88 St. Dev. = 12.57

Second test (part): Mean = 21.68 St. Dev. = 11.59

The correlation between the two sets of scores is 0.96

Taking into account apparent differences in the form means:

reliability = 0.94 st. error of measurement = 2.90

Within forms analysis:

reliability = 0.95 st. error of measurement = 2.62

Split parts analysis:

Spearman-Brown Coefficient = 0.98 and

Kuder-Richardson 20 = 0.98

Item analysis

The purpose of item analysis is to examine the contribution that each item is making to the test. Items that are identified as faulty or inefficient can be

² Statements based on the SEM tend to be less accurate when applied to people at the extremes of the distribution (the strongest and the weakest candidates). Item response theory (see below) is less susceptible to this effect.

modified or rejected. In this section of the chapter we will look first at so-called classical item analysis, before turning to a more recent development – item response theory.

Classical item analysis

This usually involves the calculation of facility values and discrimination indices, as well as an analysis of distractors in the case of multiple choice items.

Facility values

The *facility value* of an item on which only scores of zero or one can be scored is simply the proportion of test-takers that score one on it. Thus, if 100 people respond to an item and 37 give the correct response, the facility value is 0.37 (37 divided by 100). If 80 people take a test and 56 of them get an item right, the facility value is 0.70 (56 divided by 80).

What use can we make of facility values? This depends on our purpose. If we are developing a proficiency test designed to identify the top 10 percent of students for a special language course, we won't have much need for easy items, that is, items with high facility values. Those items would not discriminate between the best 10 percent and most of the other people. Ideally, for this purpose we would want a high proportion of items with a facility value not far from 0.10. If, on the other hand, we are developing a placement test which is meant to cover a wide range of abilities and place people in classes at a number of levels, we will want a wide range of facility values in our items, with no big gaps between them.

The question of facility values for items which are not scored dichotomously (that is, 1 or zero), and which may be referred to as 'partial credit' items is generally not discussed in basic texts on testing. Nevertheless, it is useful to be able to compare the difficulty of such items. What we would suggest is taking the average score on an item (i.e. total points scored on the item by all test-takers divided by the number of test-takers) and dividing that by the maximum number of points on the item. Thus, if 100 people take a five-point item and score a total of 375 points on it, the average score is 3.75 (375 divided by 100), and the facility value is 0.75 (3.75 divided by 5). The advantage of this method is that it gives the same result for zero/one items as the procedure described for them above³.

Discrimination indices

A *discrimination index* is an indicator of how well an item discriminates between weak candidates and strong candidates. The higher its discrimination index, the better the item discriminates in this way. The theoretical maximum discrimination index is 1. An item that does not discriminate at all (weak and strong candidates perform equally well on it) has a discrimination index of zero. An item that discriminates in favour

³. There are software packages which can carry out the analysis of partial credit items. There are others that can analyse items with scores that are based on rating scales.

of the weaker candidates (weaker candidates perform better than stronger candidates) – and such items are occasionally written, unfortunately – has a negative discrimination index. Discrimination is important because the more discriminating the items are, the more reliable will be the test.

Discrimination indices are typically correlation coefficients. The usual way of calculating a discrimination index is to compare performance of the candidates on the item with their performance on the test as a whole. If scores on the item (zero or one) correlate well with scores on the test, the resulting correlation coefficient will indicate good discrimination.

Strictly speaking, the correlation should be calculated between the scores made by individuals on an item and their scores on the test less their score on that item. Otherwise, scores on the item are included in scores on the test, which will exaggerate the strength of the correlation. This exaggeration is not significant when a test includes a large number of items.

Note that calculation of discrimination indices in this way assumes that, as a group, the people who do better on the whole test (or on some part of it being analysed) should do better on any particular item in it.

Look at the following discrimination indices for items in OURTEST.

ITEM 1	0.386
ITEM 2	0.601
ITEM 3	0.355
ITEM 5	0.734
ITEM 6	0.358
ITEM 7	0.434
ITEM 8	0.207
ITEM 9	0.518
ITEM 10	0.393
ITEM 11	0.590
ITEM 12	0.419
ITEM 13	0.433
ITEM 97	0.265
ITEM 98	0.469
ITEM 99	0.188
ITEM 100	0.124

The items with the greatest indices are the ones that discriminate best. The most discriminating item here, therefore, is Item 5, with an index of 0.734. The least discriminating item is Item 100, with an index of 0.124.

A natural question at this point is: What is regarded as a satisfactory discrimination index? The disappointing answer is that there is no absolute value that one can give. The important thing is the relative size of the indices. Remember that we are interested in discrimination for its effect on

reliability. The first thing we should do is look at the reliability coefficient. If there is a problem with reliability, we can look at discrimination indices to see if there are items which are not contributing enough to reliability. Any items with a negative index should be first to go. (In fact, they should be candidates for removal from the test even if the reliability coefficient is satisfactory.) After that we look for the items with the lowest positive indices. If the items themselves are clearly faulty, we should either drop them from the test (and try to replace them with better items) or we should try to improve them. A word of warning, though. An item with a low discrimination index is not necessarily faulty. Item 99 in OURTEST is a case in point. The reason for its lack of discrimination is that it is very difficult. Its facility value is only 0.022 (only two of the 186 people taking the test responded correctly). When an item is very easy or very difficult, the discrimination index is almost bound to be low. Even if an item does not discriminate well overall, we might wish to keep it in the test. If it is very easy, it might be kept because it is being used to help make the candidates feel confident at the start of the test. If it is very difficult, we may keep it because, while it does not discriminate well over all the people who took the test, it may discriminate between the strongest candidates. When OURTEST was reduced from 100 to 60 items (see above), all the items were grouped into bands according to their facility value. Then the items with the lowest discrimination indices were dropped. This is because the particular purpose of the test called for discrimination at all levels.

Where the scores of only a small number of students (say 30) are available for analysis, formal discrimination indices calculated as described above are not very meaningful. However, it is still worthwhile dividing the students into two groups – top half and bottom half (according to their scores on the complete test) – and then comparing their performance on each item. If there are items where there is no difference between the groups or where the lower group actually do better, then these items are worth scrutinising.

Analysis of distractors

Where multiple choice items are used, in addition to calculating discrimination indices and facility values, it is necessary to analyse the performance of distractors. Distractors that do not work (i.e. are chosen by very few candidates) make no contribution to test reliability. Such distractors should be replaced by better ones, or the item should be otherwise modified or dropped. However, care should be taken in the case of easy items, where there may not be many incorrect responses to be shared among the different distractors (unless a very large number of candidates have been tested).

Item response theory

Everything that has been said so far has related to classical item analysis. In recent decades new methods of analysis have been developed which

have many attractions for the test constructor. These all come under the general heading of item response theory, and the form of it so far most used in language testing is called *Rasch analysis*.

Rasch analysis begins with the assumption that items on a test have a particular difficulty attached to them, that they can be placed in order of difficulty, and that the test-taker has a fixed level of ability. Under these conditions, the idealised result of a number of candidates taking a test will be as in Table 4. The candidate with the greatest ability is 'candidate 8'; the one with the least ability is 'candidate 1'. The most difficult items are items 6 and 7; and the least difficult item is item 1.

TABLE 4: RESPONSES OF IMAGINARY CANDIDATES TO IMAGINARY ITEMS							
Candidates	Items						
	1	2	3	4	5	6	7
1	1	0	0	0	0	0	0
2	1	1	1	0	0	0	0
3	1	1	1	0	0	0	0
4	1	1	1	0	0	0	0
5	1	1	1	1	0	0	0
6	1	1	1	1	1	0	0
7	1	1	1	1	1	0	0
8	1	1	1	1	1	1	1
Total incorrect	0	1	1	4	5	7	7

(adapted from Woods and Baker 1985)

Table 4 represents an idealised model of what happens in test-taking, but we know that, even if the model is correct, people's performance will not be a perfect reflection of their ability. In the real world we would expect an individual's performance to be more like the following:

1 1 1 1 0 1 0 1 0

Rasch analysis in fact accepts such departures from the model as normal. But it does draw attention to test performance that is significantly different from what the model would predict. It identifies test-takers whose behaviour does not fit the model, and it identifies items that do not fit the model.

Here are some examples from Rasch analysis of OURTEST. It would be inappropriate (not to say impossible in the space available) to try to explain everything in the analysis. But we will just use the examples to show what it can contribute to our understanding of how items on a test are performing.

Item Number	Score	Fit
9	130	0.3252
10	160	31.6097
11	135	-3.3231
12	154	5.5788
13	156	2.2098

The first column identifies the item. The second shows how many correct responses there were on that item (out of 186). The third column shows how well the item fits the Rasch model. The higher the positive value, the less well the item fits. It can be seen that the least fitting item is Item 10, which makes us immediately suspicious of it. It's a relatively easy item (160 out of 186 candidates respond correctly); if it is misfitting, therefore, better candidates must be getting it wrong. So we look now at people that Rasch analysis identifies as misfitting. Amongst them are two who have an 'unusual' result on Item 10. The first is Person Number 10:

Person	Score	Ability	Misfit value
P10	88	3.1725	48.6729
Items with unusual result:		Item	Residual
		3	13.90
		10	29.88
		34	8.50
		60	2.54
		73	3.77
		76	2.60

We learn from the output that Person 10 had a very high score on the test (88) and performed in an unexpected way on two items in particular (Items 3 and 10 – the ones with high residuals⁴). Since these are easy items, we can conclude either that they weren't concentrating (notice that there are four other items on which there is an unusual result), or they have very surprising gaps in their knowledge, or that there is something wrong with one or both of the items.

The second person below has unusual results on eight items. The relatively small residual value for Item 10 reflects the fact that the person is of only middling ability (score 40) and so it is not so surprising that the item was responded to incorrectly.

⁴. The residual is an indication of how badly a person's performance on an item fits the Rasch model. Thus, if a candidate does very well on the test as a whole but gets a very easy item wrong, their residual for that item will be high; if they get an item of middling difficulty wrong, then the residual will be smaller. In brief, we are on the lookout for items with high residuals, because these tell us that someone's performance on that item is unexpected, i.e. doesn't fit the model.

The second is Person Number 166:

Person	Score	Ability	Misfit value
P166	40	-0.6752	4.8836
Items with unusual result:		Item	Residual
		7	4.22
		10	4.36
		61	2.77
		67	2.57
		70	4.07
		72	2.64
		81	4.64
		92	4.64

The situation so far is that we have an item that seems to misfit, and we have two people who behaved unusually on it. If we drop these two people from the analysis, the result for Item 10 is different:

ITEM 10 143 -3.7332 -3.7363

The item now fits well. When we look at the item and can find nothing wrong with it, we come to the conclusion that the problem is with the candidates, not the item. If it is thought worthwhile by the institution using the test, the two people can be followed up in an attempt to find out what is wrong.

If an item is identified as misfitting by Rasch analysis, and we cannot explain the misfit through odd performance on it by a small number of candidates, we can expect to find a problem with the item itself when we come to inspect it.

Rasch analysis assumes that what is being measured by a test is unidimensional. This parallels the assumption of classical analysis that people who do better on the test should do better on any item. Of course there may be more than one dimension to what is being learned or acquired, but this does not seem to affect the practical value of Rasch analysis any more than does classical analysis.

Another feature of Rasch analysis is that instead of giving a single standard error of measurement that has to be applied to all candidates, it gives a separate standard error for each candidate.

Thus:

Person	Ability	Standard error
P28	-5.93	0.82
P31	-3.57	0.41
P3	-0.59	0.27

Person 28 is the weakest of these three candidates (the higher the negative ability value, the weaker the person) and has the highest standard error. Person 3 is of middling ability (near zero) and has the lowest standard error. This fits with what was said in Note 2 above. We can be much more confident that Person 3's true score is close to their actual score, than we can that Person 28's true score is close to their actual score.

There is one more use of Rasch analysis to mention. Rasch analysis can be particularly helpful when we are trialling items on different groups of people. Let's say we want to trial 170 items. We believe that these are too many items to ask one group of people to respond to, so we set up two groups. The problem then is, if the two groups are not equal in ability, how can we compare the facility values of items taken by one group with the facility values of items taken by the other group? The stronger group will be putting the items on a different scale of 'easiness' from that of the weaker group. An item will be given a different facility value than it would have had if it had been taken by the other group.

The answer to this problem is to use what are called *anchor items*. These are items, preferably ones that are known to be 'good', which both groups respond to. So in the situation referred to, 30 items could be anchors. The remaining 140 items would be split into two sets, so that each group took a total of 100 items. Once the items have been administered and scored, Rasch analysis has the ability to use the common anchor items to put all of the other items on the same scale. With the increased use of item banks (Appendix 1), this is a particularly valuable feature.

There is one last thing to say about item analysis of the kind we have described. As we hope to have shown, both classical analysis and Rasch analysis have contributions to make to the development of better tests. They should be seen as complementary, not in opposition with one to be chosen over the other.

The analysis of criterion-referenced tests

The analysis of criterion-referenced tests differs somewhat from that of norm-referenced tests.

The descriptive statistics (mean, standard deviation, etc.) which we have presented in this chapter may also be calculated for criterion-referenced tests.

We have already shown (Chapter 5) how to calculate a measure of decision consistency, a criterion-referenced equivalent to the reliability coefficient of norm-referenced tests.

Turning to item analysis, facility values are calculated in the same way as those for norm-referenced tests.

The calculation of *discrimination indices*, however, is different. In criterion-referenced tests, we are interested in how well an item discriminates between those who have reached the criterial level on the whole test (Achievers), and those who have not (Non-Achievers).

1. We begin with the separation of candidates into Achievers (A) and Non-Achievers (N-A). Normally, the number in each of the two groups will be different.

For example, 78 out of 112 candidates may reach the criterion level: so there are 78 in the A group, and 34 in the N-A group.

2. Then for each item, we calculate the proportion in each group who were successful on that item. For example, if 72 in the A group are successful on an item, and 12 in the N-A group are successful, then the proportions are 0.92 and 0.35.
3. The next step is to subtract the proportion of the N-A group from the proportion of the A group. This gives the discrimination index. So, continuing our example, $0.92 - 0.35 = 0.57$.

We should look closely at any item with a very low discrimination index and satisfy ourselves that there is a good reason for it (e.g. it is an easy item that you would expect few in the N-A group to respond to incorrectly). A negative discrimination index, indicating that the N-A group did better than the A group, suggests that there is a problem with the item.

Final word

This chapter on the statistical analysis of tests will not have pleased everyone. For many readers statistics will have little, if any, appeal. Other readers may be frustrated that the treatment of the subject has been so sketchy. Our only hope is that there will at least be some people who find it sufficiently interesting and potentially useful to them that they will go on to experiment with statistics in their language testing, and to study the subject in greater depth.



FURTHER READING

For the use of statistics in language studies more generally, see Woods et al. (1986). For an introduction to item response theory, see Woods and Baker (1985). For a much fuller treatment, see Chapters 5–9 of McNamara (1996). Brown (2002, 2009) shows how some test statistics can be carried out using a spreadsheet.

Online resources

ITEMAN is a basic item analysis package.

Appendix 1 Item banking

A supply of good items for use in the construction of future tests is clearly a valuable asset. In the past, such items were often typed onto record cards and stored in box files. These days, collections of items, known as item banks, are held on computers.

With each item is stored information that has been derived from statistical analysis of the kind described in Chapter 19. The information includes:

1. A number of identifying criteria, relating to such things as its content, class level, stage in the syllabus or coursebook, the testing technique used, and number of points.
2. Correct response(s) and scoring instructions.
3. Measurement information on the item, such as difficulty level and discrimination index, which has been obtained through previous trialling.
4. Notes on the item (when written, when used, etc.).

Once they have access to an item bank, test constructors simply choose from it the items that they need for a test. They do this by entering into the computer details of the kinds of item they need. They might begin, for example, by asking for receptive vocabulary items which have a facility value between 0.4 and 0.6, and which relate to third-year study at their institution. The computer will immediately present them with all the items in the bank that meet these criteria, and they are given the opportunity to 'browse' through these, choosing those items that they decide to include in the test. Once they have chosen all the items they need for the test, and have provided details such as the test title and rubrics, the computer provides a printed version of the test.

There are a number of benefits to be had from item banks:

1. Once the bank is constructed, there is a considerable saving of effort. Tests do not have to be constructed over and over again from scratch.
2. Since the trialling of the items (which makes use of anchor items) is carried out before they are entered into the bank, the quality of tests that use them will almost certainly be higher than those made up of untried items.
3. The psychometric information on items gathered during trialling means that the measurement qualities (including level of difficulty) of tests made up of these items can be predicted (before the test is

taken) with greater accuracy than when predictions are made on the basis of test constructors' judgements. This in turn means that a test constructed in one year can be made to have the same difficulty as tests set in previous years, with implications for the maintenance of standards, fairness and the evaluation of teaching.

The development of an item bank follows very much the procedures as those for the development of a test. The only differences are that the specifications have to be for a *bank*, not a test; and the trialling process – making use of anchor items – is absolutely essential.

Item banks are now regarded as indispensable to serious testing organisations. With the advent of powerful but inexpensive computers, item banks have become an attractive possibility for all serious testers who are prepared to put in the necessary initial effort.

Relatively simple item banks can be constructed using spreadsheet software. But there are dozens of commercially available software programs specifically designed for item banking. Readers are warned, however, that they should take great care to ensure that any program they think of buying will cater fully for their needs, since many of them are quite restrictive in the kinds of items which they can incorporate. They should also be sure that the program allows them to enter all the information about the items that they want to include.

Appendix 2 Checklist for teachers choosing tests for their students

- Be clear about what you want the test for (see Chapter 2).
 - Do you simply want to measure your student's ability (proficiency test)?
 - Do you want to measure how successfully students have achieved the objectives of the course (achievement test)?
 - Do you want to measure students' abilities in order to place them in appropriate levels (placement test)?
 - Do you want to identify the strengths and weaknesses of your students in order to inform future teaching (diagnostic test)?
- Be clear on what your students want the test for. For example:
 - to apply to a university
 - to become a member of a professional body
 - to find out what their level is in a particular skill
 - to receive feedback
- Consider whether the test will be taken externally by students or used in-house. If the test is to be used in-house, consider the practical implications. How will the test be scored? Is the test compatible with the school's hardware?
- Search for information online. The *languagetesting.info* website lists major test providers as well as links to a huge number of useful resources.
- Having considered the steps described above, make a shortlist of possible tests for your students. For each of these tests, look for:
 - Evidence of acceptance by institutions (if that is what you and your students are looking for).
 - Evidence of validity. A respectable test should present evidence of validity. Chapter 4 will help you evaluate this.
 - Evidence of reliability. Refer to Chapter 5 to help you evaluate this.
 - Likely backwash (see Chapter 6). Although backwash is irrelevant when students *have to* take a test in order to gain a qualification, it should be considered when teachers are choosing a test for other reasons.

- Availability of a test handbook. Handbooks help students and teachers to become familiar with the structure of the test and the nature of the items.
- Availability of practice materials. Larger organisations provide official practice materials either online or as published books. There are also countless unofficial practice materials online but these can vary in quality and usefulness. Teachers should evaluate the usefulness of these unofficial materials and select accordingly.

Appendix 3 The secrets of happiness

For questions 1 and 2, the sentences in the article which give you the answers have been underlined. Read the questions and the underlined sentences. Then choose the answer (A, B, C or D) which you think fits best according to the underlined sentences.

- 1 What does *this* in line 6 refer to?
A the writer's decision to study psychology
B the writer's interest in happiness
C the writer's observations of adults
D the writer's unhappy childhood
- 2 What sort of people did the writer choose to concentrate on at the start of his career?
A People who were clearly happier
B People with more freedom
C People whose main aim in life was not making money
D People whose objective was to become richer

Now, for questions 3–6, choose the answer (A, B, C or D) which you think fits best according to the text.

- 3 The 'experience sampling method' showed in general that
A creative people are happier than other people.
B uncreative people are just as happy as creative people.
C people's happiness depends on who they are with.
D people are happier when they are very focused on an activity.
- 4 *that dividing line* in line 47 refers to a division between
A living more comfortably and less comfortably.
B poor countries and rich countries.
C happy people and unhappy people.
D millionaires and poor people.
- 5 According to the writer, people concentrate more when they are doing
A something which they find enjoyable.
B something which they find difficult but possible.
C something which they find quite easy.
D many things at the same time.
- 6 What impression do you have of the writer of the text?
A He has become happier by studying happiness.
B He has been unhappy most of his life.
C He has always been a happy person.
D He has only been happy for short times.

Bibliography

- Adams, M. L. and J. R. Frith (Eds.). 1979. *Testing Kit*. Washington DC: Foreign Service Institute.
- AERA. 1999. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.
- Alderson, J. C. 1990a. Testing reading comprehension skills (Part one). *Reading in a Foreign Language* 6: 425–438.
- Alderson, J. C. 1990b. Testing reading comprehension skills (Part two). *Reading in a Foreign Language* 6: 465–503.
- Alderson, J. C. 1995. Response to Lumley. *Language Testing* 12: 121–125.
- Alderson, J. C. 2000. *Assessing reading*. Cambridge: Cambridge University Press.
- Alderson, J. C. 2005. *Diagnosing foreign language proficiency: the interface between learning and assessment*. London: Continuum.
- Alderson, J. C. 2009. Review of Test of English as a Foreign Language: internet-based test (TOEFL iBT). *Language Testing* 26: 621–631.
- Alderson, J. C. and G. Buck. 1993. Standards in testing: a study of the practice of UK examination boards in EFL/ESL testing. *Language Testing* 10: 1–26.
- Alderson, J. C. and L. Hamp-Lyons. 1996. TOEFL preparation courses: a study of washback. *Language Testing* 13: 280–297.
- Alderson, J. C. and A. Hughes (Eds.). 1981. Issues in language testing. *ELT Documents* 111. London: The British Council.
- Alderson, J. C., and B. Kremmel. 2013. Re-examining the content validation of a grammar test: the (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing* 30: 535–556.
- Alderson, J. C. and D. Wall. 1993. Does Washback exist? *Applied Linguistics* 14: 115–129.
- Alderson, J. C., Clapham, C. and D. Wall. 1995. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.
- Alderson, J. C., Percicsich R. and G. Szabo. 2000. Sequencing as an item type. *Language Testing* 17: 421–447.
- Alderson, J. C., Brunfaut, T. and L. Harding. 2015. Towards a theory of diagnosis in second and foreign language assessment: insights from professional practice across diverse fields. *Applied Linguistics* 36: 236–260.
- Allan, A. 1992. Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing* 9: 101–122.

- Al Mahrooqi, R. and C. Denman. 2018. Alternative Assessment. In Liontas, J. I. 2018. *The TESOL encyclopedia of English language teaching*. New Jersey: John Wiley and Sons.
- Amini, M. and N. Ibrahim-González. 2012. The Washback Effect of Cloze and Multiple-Choice Tests on Vocabulary Acquisition. *Language in India*, 12.
- Anastasi, A. and S. Urbina. 1997. *Psychological Testing* (7th edition). Upper Saddle River, N.J: Prentice Hall.
- Anderson, N. J. 2012. Student involvement in assessment: healthy self-assessment and effective peer assessment. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyloff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Arnold, J. 2000. Seeing through listening comprehension anxiety. *TESOL Quarterly* 34: 777–786.
- Aryadoust, V. and L. Zhang. 2016. Fitting the mixed Rasch model to a reading comprehension test: Exploring individual difference profiles in L2 reading. *Language Testing* 33: 529–553.
- Babaii, E., Taghaddomi, S. and R. Pashmforoosh. 2016. Speaking self-assessment: mismatches between learners' and teachers' criteria. *Language Testing* 33: 411–437.
- Bachman, L. F. and A. D. Cohen (Eds.). 1998. *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.
- Bachman, L. F. and A. S. Palmer. 1996. *Language testing in practice*. Oxford: OUP.
- Bachman, L. F. and A. S. Palmer. 2010. *Language assessment in practice: developing assessments and justifying their use in the real world*. Oxford: Oxford University Press.
- Bailey, K. M. 1996. Working for Washback: a review of the Washback concept in language testing. *Language Testing* 13: 257–279.
- Batty, A. O. 2015. A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing* 32: 3–20.
- Benzehra, R. 2018. Multiple-measures assessment. In Liontas, J. I. 2018. *The TESOL encyclopedia of English language teaching*. New Jersey: John Wiley and Sons.
- Bernstein, J., Van Moere, A. and J. Cheng. 2010. Validating automated speaking tests. *Language Testing* 27: 355–377.
- Bouwer, R., Beguin, A., Sanders, T. and H. Van den Bergh. 2014. Effect of genre on the generalizability of writing scores. *Language Testing* 3: 83–100.
- Boyd, K. and A. Davies. 2002. Doctors' orders for language testers: the origin and purpose of ethical codes. *Language Testing* 3: 296–322.

- Bradshaw, J. 1990. Test-takers' reactions to a placement test. *Language Testing* 7: 13–30.
- Brett, P. and G. Motteram (Eds.). 2000. *A special interest in computers: learning and teaching with information and communications technologies*. Whitstable: IATEFL.
- Brook-Hart, G. 2014. *Complete First Second Edition Student's Book*. Cambridge: Cambridge University Press.
- Brown, A. 2003. Interviewer variation and the co-construction of speaking proficiency. *Language Testing* 20: 1–25.
- Brown, A. 2012. Ethics in language testing and assessment. In Coombe C., Davidson, P., O'Sullivan, B. and S. Stoyhoff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Brown, J. D. 1990. Short-cut estimates of criterion-referenced test consistency. *Language Testing* 7: 77–97.
- Brown, J. D. 1993. What are the characteristics of *natural* cloze tests? *Language Testing* 10: 93–115.
- Brown, J. D. 2002. Statistics Corner. Questions and answers about language testing statistics: Distractor efficiency analysis on a spreadsheet. *Shiken: JALT Testing & Evaluation SIG Newsletter* 63: 20–23. (also available online)
- Brown, J. D. 2009. Using a spreadsheet program to record, organize, analyze, and understand your classroom assessments. In Coombe, C., Davidson, P. and D. Lloyd (Eds.). 2009. *The Fundamentals of language assessment: A practical guide for teachers* (2nd edition). Dubai, UAE: TESOL Arabia Publications.
- Brown, J. D. 2012. Choosing the right type of assessment. In Coombe C., Davidson, P., O'Sullivan, B. and S. Stoyhoff. (Eds.). *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Brown, J. D. and T. Hudson. 1998. The alternatives in language assessment. *TESOL Quarterly* 32: 653–675.
- Brown, J. D. and T. Hudson. 2002. *Criterion-referenced language testing*. Cambridge: Cambridge University Press.
- Buck, G. 1991. The testing of listening comprehension: an introspective study. *Language Testing* 8: 67–91.
- Buck, G. 2001. *Assessing listening*. Cambridge: Cambridge University Press.
- Buck, G. and K. Tatsuoaka. 1998. Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing* 15: 119–157.
- Butler, Y. G. and J. Lee. 2010. The effects of self-assessment among young learners of English. *Language Testing* 27: 5–31.

- Bygate, M. 1987. *Speaking*. Oxford: Oxford University Press.
- Cai, H. 2013. Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing* 30: 177–199.
- Cameron, L. 2001. *Teaching language to young learners*. Cambridge: Cambridge University Press.
- Canale, M. and M. Swain. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1: 1–47.
- Carey, M. D., Mannell, R. H. and P. K. Dunn. 2011. Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing* 28: 201–219.
- Carpenter, K., Fujii, N. and H. Kataoka. 1995. An oral interview procedure for assessing second language abilities in children. *Language Testing* 12: 157–181.
- Carroll, J. B. 1961. Fundamental considerations in testing for English language proficiency of foreign students. In Allen, H. B. and R. N. Campbell (Eds.). 1972. *Teaching English as a second language: a book of readings*. New York: McGraw-Hill.
- Chalhoub-Deville, M. (Ed.) 1999. *Issues in computer adaptive testing of reading proficiency: selected papers*. Cambridge: Cambridge University Press.
- Chalhoub-Deville, M. and C. Deville. 1999. Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics* 19: 273–299.
- Chapelle, C. A. and R. G. Abraham. 1990. Cloze method: what difference does it make? *Language Testing* 7: 121–146.
- Chapelle, C. A. and D. Douglas. 2006. *Assessing language through computer technology*. Cambridge: Cambridge University Press.
- Chapelle, C. A., Enright, M. K. and J. M. Jamieson. 2008. *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.
- Chapelle, C. A., Chung, Y. R., Hegelheimer, V., Pendar, N. and J. Xu. 2010. Towards a computer-delivered test of productive grammatical ability. *Language Testing* 27: 443–469.
- Cheng, L. 2005. Changing language teaching through language testing: a washback study. *Studies in Language Testing* 21. Cambridge: Cambridge University Press.
- Cheng, L. and A. Curtis. 2012. Test impact and washback: implications for teaching and learning. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyanoff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.

- Cheng, L. and J. Fox. 2017. *Assessment in the language classroom*. Oxford: Macmillan (Red Globe Press).
- Cheng, L., Andrews, S. and Y. Yu. 2011. Impact and consequences of school based assessment (SBA): students' and parents' views of SBA in Hong Kong. *Language Testing* 28: 221–249.
- Cheng, L., Watanabe, J. and A. Curtis (Eds.). 2004. *Washback in language testing: research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum.
- Cheng, W. and M. Warren. 2005. Peer assessment of language proficiency. *Language Testing* 22: 93–121.
- Chester, M. D. 2005. *Multiple Measures and High-Stakes Decisions: A Framework for Combining Measures*. Wiley Online Library. (Accessed online: <https://doi.org/10.1111/j.1745-3992.2003.tb00126.x>)
- Choi, I. C. 2008. The impact of EFL testing on EFL education in Korea. *Language Testing* 25: 39–62.
- Clapham, C. and D. Corson (Eds.). 1997. *Encyclopaedia of Language and Education. Volume 7: Language testing and assessment*. Amsterdam: Kluwer Academic Publishers.
- Cohen, A. D. 1984. On taking language tests: What the students report. *Language Testing* 1: 70–81.
- Cohen, A. D. 2012. Test-taking strategies. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyhoff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Collins Cobuild. 1992. *English Usage*. London: Harper Collins.
- Coombe, C. 2010. Assessing Foreign/Second Language Writing Ability. *Education, Business and Society: Contemporary Middle Eastern Issues* 3: 178–187.
- Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyhoff (Eds.). 2012a. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Coombe, C., Purmenschky, K. and P. Davidson. 2012b. *Alternative assessment in language education*. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyhoff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Coombe, C., Troudi, S. and M. Al-Hamly. 2012c. Foreign and second language teacher assessment literacy: issues, challenges and recommendations. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyhoff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Council of Europe. 2001. *Common European framework of references for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

- Criper, C. and A. Davies. 1988. *ELTS validation project report*. Cambridge: The British Council and Cambridge Local Examinations Syndicate.
- Crystal, D. and D. Davy. 1975. *Advanced conversational English*. London: Longman.
- Cumming, A. and R. Berwick (Eds.). 1996. *Validation in language testing*. Clevedon: Multilingual Matters.
- Cumming, A., Grant, L., Mulcahy-Ernt, P. and D. E. Powers. 2004. A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing* 21: 107–145.
- Currie, M. and T. Chiraramanee. 2010. The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing* 27: 471–491.
- Curtis, A. 2018. Portfolios. In Leontis, J. I. 2018. *The TESOL encyclopedia of English language teaching*. New Jersey: John Wiley and Sons.
- Dávid, G. 2007. Investigating the performance of alternative types of grammar items. *Language Testing* 24: 65–97.
- Davidson, F. 2000. Review of *Standards for educational and psychological testing*. *Language Testing* 17: 457–462.
- Davidson, F. and G. Fulcher. 2012. Developing test specifications for language assessment. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyonoff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Davies, A. (Ed.). 1968. *Language testing symposium: a psycholinguistic perspective*. Oxford: Oxford University Press.
- Davies, A. 1988. *Communicative language testing*. In Hughes, A. (Ed.). 1988b. *Testing English for university study. ELT Documents 127*. Oxford: Modern English Press.
- Davies, A. 2010. Test fairness: a response. *Language Testing* 27: 171–176.
- Drackert, A., and A. Timukova. 2020. What does the analysis of C-test gaps tell us about the construct of a C-test? A comparison of foreign and heritage language learners' performance. *Language Testing* 37: 107–132.
- Ebel, R. L. 1978. The case for norm-referenced measurements. *Educational Researcher* 7: 3–5.
- Eckes, T. 2008. Rater types in writing performance assessments: a classification approach to rater variability. *Language Testing* 25: 155–185.
- Elder, C., Barkhuizen, G., Knoch, U., and J. von Randow. 2007. Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing* 24: 37–64.
- Farhady, H. and M. N. Keramati. 1996. A text-driven method for the deletion procedure in cloze passages. *Language Testing* 13: 191–207.

- Feldt, L. S. and R. L. Brennan. 1989. Reliability. In Linn, R. L. (Ed.). 1989. *Educational measurement* (7th edition). New York: Macmillan.
- Ferris, D. R. 2002. *Treatment of error in second language student writing*. Ann Arbor: University of Michigan Press.
- Field, J. 2019. *Rethinking the Second Language Listening Test: From Theory to Practice*. British Council Monographs on Modern Language Testing 2. Sheffield: Equinox Publishing.
- Fitzpatrick, T. and J. Clenton. 2010. The challenge of validation: assessing the performance of a test of productive vocabulary. *Language Testing* 27: 537–554.
- Fox, J. 2004. Test decisions over time: tracking validity. *Language Testing* 21: 437–465.
- Freedle, R. and I. Kostin. 1993. The prediction of TOEFL reading item difficulty: implications for construct validity. *Language Testing* 10: 133–170.
- Freedle, R. and I. Kostin. 1999. Does the text matter in a multiple-choice test of comprehension: The case for the construct validity of TOEFL's minitalks. *Language Testing* 16: 2–32.
- Fulcher, G. 1997. An English language placement test: issues in reliability and validity. *Language Testing* 14: 113–138.
- Fulcher, G. 2000. *Computers in language testing*. In Brett, P. and G. Motteram (Eds.). 2000. *A special interest in computers: learning and teaching with information and communications technologies*. Whitstable: IATEFL.
- Fulcher, G. 2003. *Testing second language speaking*. Harlow: Pearson Longman.
- Genesee, F. and J. Upshur. 1996. *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.
- Ginther, A. 2002. Context and content visuals and performance on listening comprehension stimuli. *Language Testing* 19: 133–167.
- Gipps, C. 1990. *Assessment: A teacher's guide to the issues*. London: Hodder and Stoughton.
- Godshalk, F. L., Swineford, F. and W. E. Coffman. 1966. *The measurement of writing ability*. New York: College Entrance Examination Board.
- Green, A. B. 2007. IELTS washback in context: preparation for academic writing in higher education. *Studies in Language Testing* 25. Cambridge: Cambridge University Press.
- Green, A. B. 2012. Placement testing. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyloff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Green, A. B. and C. J. Weir. 2004. Can placement tests inform instructional decisions? *Language Testing* 21: 467–494.

- Green, A. B., Ünalı, A. and C. Weir. 2010. Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing* 27(2): 191–211.
- Hale, G. A. and R. Courtney. 1994. The effects of note-taking on listening comprehension in the Test of English as a Foreign Language. *Language Testing* 11: 29–47.
- Hamp-Lyons, L. 1997a. Washback, impact and validity: ethical concerns. *Language Testing* 14: 295–303.
- Hamp-Lyons, L. 1997b. Ethical test preparation practice: the case of the TOEFL. *TESOL Quarterly* 32: 329–337.
- Hamp-Lyons, L. 1999. The author responds. *TESOL Quarterly* 33: 270–274.
- Harding, L. 2012. Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing* 29: 163–180.
- Harris, D. P. 1968. *Testing English as a second language*. New York: McGraw-Hill.
- Harsch, C. and J. Hartig. 2016. Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing* 33: 555–575.
- Hasselgren, A. 1999. *Kartlegging av Kommunikativ Kompetanse i Engelsk (Testing of Communicative Ability in English)*. Oslo: Nasjonalt læremid-delsenter.
- Hasselgreen, A. and G. Caudwell. 2016. *Assessing the language of young learners*. Sheffield: Equinox Publishing.
- Heaton, J. B. 1975. *Writing English language tests*. London: Longman.
- Hubley, N. J. 2012. Assessing reading. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyloff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Hudson, T. and B. Lynch. 1984. A criterion-referenced approach to ESL achievement testing. *Language Testing* 1: 171–201.
- Hughes, A. 1981. Conversational cloze as a measure of oral ability. *English Language Teaching Journal* 35: 161–168.
- Hughes, A. 1986. A pragmatic approach to criterion-referenced foreign language testing. In Portal, M. (Ed.). 1986. *Innovations in language testing*. Windsor: NFERNelson.
- Hughes, A. 1988a. Introducing a needs-based test of English for study in an English medium university in Turkey. In Hughes, A. 1988b.
- Hughes, A. (Ed.). 1988b. Testing English for university study. *ELT Documents* 127. Oxford: Modern English Press.
- Hughes, A. 1993. Backwash and TOEFL 2000. Unpublished paper commissioned by Educational Testing Services.

- Hughes, A. 2011. *The pursuit of truth*. Kibworth Beauchamp: Matador.
- Hughes, A. and D. Porter (Eds.). 1983. *Current developments in language testing*. London: Academic Press.
- Hughes, A. Gülçur, L., P., Gürel, P. and T. McCombie. 1987. The new Bogaziçi University English Language Proficiency Test. In Bozok, S. and A. Hughes. *Proceedings of the seminar, Testing English beyond the high school*. Istanbul: Bogaziçi University Publications.
- Hughes, A., Trudgill, P. and D. Watt. 2012. *English accents and dialects: an introduction to social and regional varieties of British English* (5th edition). London: Hodder Education.
- Hughes, A., Porter, D. and C. J. Weir (Eds.). 1988. *Validating the ELTS test: a critical review*. Cambridge: The British Council and University of Cambridge Local Examinations Syndicate.
- Hughes, A., Porter, D. and C. J. Weir. 1996. *ARELS Placement Test* [Written]. London: ARELS.
- Hughes, J. and F. Scott-Barrett. 2017. *C21: English for the 21st century: Level 5*. Reading: Garnet Publishing Limited.
- Huhta, A., Kalaja, P. and A. Pitkänen-Huhta. 2006. Discursive construction of a high-stakes test: the many faces of a test-taker. *Language Testing* 23: 326–350.
- Hyland, K. and F. Hyland. 2006. *Feedback in second language writing: contexts and issues*. Cambridge: Cambridge University Press.
- In'nami, Y. and R. Koizumi. 2009. A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing* 26: 219–244.
- Ioannou-Georgiou, P. 2003. *Assessing young language learners*. Oxford: Oxford University Press.
- Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfield, V. F. and J. B. Hughey. 1981. *Testing ESL composition: a practical approach*. Rowley, MA: Newbury House.
- Jafarpur, A. 1995. Is C-testing superior to cloze? *Language Testing* 12: 194–216.
- Jang, E. E. 2009. Cognitive diagnostic assessment of L2 reading comprehension ability: validity arguments for Fusion Model application to *LanguEdge* assessment. *Language Testing* 26: 31–73.
- Jennings, M., Fox, J., Graves, B. and E. Shohamy. 1999. The test-taker's choice: an investigation of the effect of topic on language test-performance. *Language Testing* 16: 426–456.
- Johnson, D. M. and L. Hamp-Lyons. 1995. Rating non-native writing: the trouble with holistic scoring. *TESOL Quarterly* 29: 759–762.
- Kane, M. 2010. Validity and fairness. *Language Testing* 27: 177–182.

- Kane, M. 2011. Review of Bachman and Palmer (2010). *Language Testing* 28: 581–587.
- Katz, A. 2012. *Linking assessment with instructional aims*. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyanoff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Kim, H. and C. Elder. 2015. Interrogating the construct of aviation English: feedback from test takers in Korea. *Language Testing* 32: 129–149.
- Klein-Braley, C. 1985. A cloze-up on the C-Test: a study in the construct validation of authentic tests. *Language Testing* 2: 76–104.
- Klein-Braley, C. 1997. C-Tests in the context of reduced redundancy testing: an appraisal. *Language Testing* 14: 47–84.
- Klein-Braley, C. and U. Raatz. 1984. A survey of research on the C-Test. *Language Testing* 1: 131–146.
- Knoch, U. 2009. Diagnostic assessment of writing: a comparison of two rating scales. *Language Testing* 26: 275–304.
- Kobayashi, M. 2002. Method effects on reading comprehension test performance: text organization and response format. *Language Testing* 19: 193–220.
- Kokhan, K. 2013. An argument against using standardized test scores for placement of international undergraduate students in English as a Second Language (ESL) courses. *Language Testing* 30: 467–489.
- Krekeler, C. 2006. Language for special academic purposes (LSAP) testing: the effect of background knowledge revisited. *Language Testing* 23: 99–130.
- Kunnan, A. J. 2010. Test fairness and Toulmin's argument structure. *Language Testing* 27: 183–189.
- Lado, R. 1961. *Language testing*. London: Longman.
- Lado, R. 1986. Analysis of native speaker performance on a cloze test. *Language Testing* 3: 130–146.
- Lam, R. 2015. Language assessment training in Hong Kong: implications for language assessment literacy. *Language Testing* 32: 169–198.
- Laufer, B. and Z. Goldstein. 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning* 54: 399–436.
- Lazaraton, A. 1996. Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing* 13: 151–172.
- Lee-Ellis, S. 2009. The development and validation of a Korean C-Test using Rasch analysis. *Language Testing* 26: 245–274.
- Lewkowicz, J. A. 2000. Authenticity in language testing: some outstanding questions. *Language Testing* 17: 43–64.

- Linn, R. L. (Ed.). 1989. *Educational measurement* (7th edition). New York: Macmillan.
- Liontas, J. I. (Ed.). 2018. *The TESOL encyclopedia of English language teaching*. New Jersey: John Wiley and Sons.
- Lumley, T. 1993. The notion of subskills in reading comprehension tests: an EAP example. *Language Testing* 10: 211–234.
- Lumley, T. 1995. Reply to Alderson's response. *Language Testing* 12: 125–130.
- Lumley, T. and T. F. McNamara. 1995. Rater characteristics and rater bias: implications for training. *Language Testing* 12: 54–71.
- Luoma, S. and M. Tärnänen. 2003. Creating a self-rating instrument for second language writing: from idea to implementation. *Language Testing* 20: 440–465.
- Luxia, Q. 2005. Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing* 22: 142–173.
- Matsuno, S. 2009. Self-, peer-, and teacher assessments in Japanese university EFL writing classrooms. *Language Testing* 26: 75–100.
- McKay, P. 2006. *Assessing young language learners*. Cambridge: Cambridge University Press.
- McNamara, T. 1996. *Measuring second language performance*. London: Longman.
- McNamara, T. and C. Roever. 2006. *Language testing: the social dimension*. Oxford: Blackwell Publishing.
- Messick, S. 1989. *Validity*. In Linn, R. L. (Ed.). 1989. *Educational measurement* (7th edition). New York: Macmillan.
- Messick, S. 1996. Validity and washback in language testing. *Language Testing* 13: 241–256.
- Midraj, J. 2018. Self-assessment. In Liontas, J. I. (Ed.). 2018. *The TESOL encyclopedia of English language teaching*. New Jersey: John Wiley and Sons.
- Morrow, J. 1979. Communicative language testing: revolution or evolution? In Brumfit, C. J. and K. Johnson. *The communicative approach to language teaching*. Oxford: Oxford University Press. Reprinted in Alderson, J. C. and A. Hughes (Eds.). 1981. *Issues in language testing. ELT Documents* 111. London: The British Council.
- Morrow, K. 1986. The evaluation of tests of communicative performance. In Portal, M. (Ed.). 1986. *Innovations in language testing*. Windsor: NFERNelson.
- Morrow, K. 2012. Communicative language testing. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyhoff (Eds.). 2012. Cambridge: Cambridge University Press.
- Muñoz, A. P. and E. Álvarez 2010. Washback of an oral assessment system in the EFL classroom. *Language Testing* 27: 33–49.

- Muñoz, A.P. and M. E. Álvarez. 2010. Washback of an oral assessment system in the EFL classroom. *Language Testing* 27: 33–49.
- Nakatsuhara, F. 2011. Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing* 28: 483–508.
- Nitko, A. J. 2001. *Educational assessment of students* (3rd edition). Upper Saddle River, New Jersey: Prentice Hall.
- Nixon, C. and M. Tomlinson. 2018. *Power Up Level 1 Activity Book*. Cambridge: Cambridge University Press.
- North, B. and G. Schneider. 1998. Scaling descriptors for language proficiency scales. *Language Testing* 15: 217–263.
- Ockey, G. J. 2007. Construct implications of including still image or video in computer-based listening tests. *Language Testing* 24: 517–537.
- Ockey, G. J. 2009. The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing* 26: 161–186.
- Ockey, G. and E. Wagner. 2018. *Assessing L2 listening*. Philadelphia: John Benjamins Publishing Company.
- Oller, J. W. 1979. *Language tests at school: a pragmatic approach*. London: Longman.
- Oller, J. W. and C. A. Conrad. 1971. The cloze technique and ESL proficiency. *Language Learning* 21: 183–194.
- O'Loughlin, K. 2002. The impact of gender in oral proficiency testing. *Language Testing* 19(2): 169–192.
- O'Loughlin, K. J. 2001. *The equivalence of direct and semi-direct speaking tests*. Cambridge: Cambridge University Press.
- O'Sullivan, B. 2012a. Assessing speaking. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyanoff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- O'Sullivan, B. 2012b. The assessment development process. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyanoff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Pilliner, A. 1968. *Subjective and objective testing*. In Davies, A. (Ed.). 1968. *Language testing symposium: a psycholinguistic perspective*. Oxford: Oxford University Press.
- Pollitt, A. 2012. The method of Adaptive Comparative Judgement. *Assessment in Education, Policy and Practice* 19: 281–300.
- Popham, W. J. 1978. The case for criterion-referenced measurements. *Educational Researcher* 7: 6–10.
- Portal, M. (Ed.). 1986. *Innovations in language testing*. Windsor: NFER-Nelson.

- Qian, D. D. and M. Schedl. 2004. Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing* 21: 28–52.
- Quirke, P. 2018a. Observations. In Liontas, J. I. (Ed.). 2018. *The TESOL encyclopedia of English language teaching*. New Jersey: John Wiley and Sons.
- Quirke, P. 2018b. Journals. In Liontas, J. I. (Ed.). 2018. *The TESOL encyclopedia of English language teaching*. New Jersey: John Wiley and Sons.
- Raven, J. 1991. *The tragic illusion: Educational testing*. Unionville, NY: Trillium Press and Oxford: Oxford Psychologists Press.
- Read, J. 2000. *Assessing vocabulary*. Cambridge: Cambridge University Press.
- Read, J. 2007. Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies* 7: 105–126.
- Read, J. and C. A. Chapelle. 2001. A framework for second language vocabulary assessment. *Language Testing* 18: 1–32.
- Rea-Dickens, P. 1997. So, why do we need relationships with stakeholders in language testing? A view from the UK. *Language Testing* 14: 304–314.
- Rea-Dickens, P. and S. Rixon. 1997. The assessment of young learners of English as a foreign language. In Clapham, C. and D. Corson (Eds.). 1997. *Encyclopaedia of Language and Education. Volume 7: Language testing and assessment*. Amsterdam: Kluwer Academic Publishers.
- Riazi, M. 2014. *Research Note: Concurrent and predictive validity of Pearson Test of English Academic (PTE Academic)*. (Available online at: https://pearsonpte.com/wp-content/uploads/2014/07/Riazi_M_2014.pdf)
- Rimmer, W. 2006. Measuring grammatical complexity: The Gordian knot. *Language Testing* 23: 497–519.
- Rixon, S. 2013. *Survey of policy and practice in primary English language teaching worldwide*. London: British Council.
- Roever, C. 2011. Testing of second language pragmatics: Past and future. *Language Testing* 28: 463–481.
- Roever, C. and G. Kasper. 2018. Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing* 35: 331–355.
- Rupp, A. A., Ferne, T. and H. Choi. 2006. How assessing reading comprehension with multi-choice questions shapes the construct: A cognitive processing perspective. *Language Testing* 23: 441–474.
- Ryan, K. 2011. Book reviews. *Language Testing* 28: 298–304.
- Sadhvani, P. and D. Sheetz. 2018. Presentations. In Liontas, J. I. (Ed.). 2018. *The TESOL encyclopedia of English language teaching*. New Jersey: John Wiley and Sons.

- Saif, S. 2006. Aiming for positive washback: a case study of international teaching assistants. *Language Testing* 23: 1–34.
- Saito, H. 2008. EFL Classroom peer assessment: training effects on rating and commenting. *Language Testing* 25: 553–581.
- Scott, M. L., Stansfield, C. W. and D. M. Kenyon. 1996. Examining validity in a performance test: the listening summary translation exam (LSTE) – Spanish version. *Language Testing* 13: 83–109.
- Shaw, S. D. and C. J. Weir. 2007. *Examining writing: research and practice in assessing second language writing*. Cambridge: Cambridge University Press.
- Shehadeh, A. 2012. *Task-based assessment: components, development, and implementation*. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyonoff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Sherman, J. 1997. The effect of question preview in listening comprehension tests. *Language Testing* 14: 185–213.
- Shizuka, T., Takeuchi, O., Yashima, T. and K. Yoshizawa. 2006. A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing* 23: 35–57.
- Shohamy, E. 2001. *The power of tests: a critical perspective on the uses of language tests*. Abingdon: Routledge.
- Shohamy, E. and O. Inbar. 1991. Validation of listening comprehension tests: the effect of text and question type. *Language Testing* 8: 23–40.
- Shohamy, E. and T. McNamara. 2009. Language Tests for Citizenship, Immigration, and Asylum. *Language Assessment Quarterly* 6: 1–5.
- Shohamy, E., Donitsa-Schmidt, S. and I. Ferman. 1996. Test impact revisited: Washback effect over time. *Language Testing* 13: 298–317.
- Skehan, P. 1984. Issues in the testing of English for specific purposes. *Language Testing* 1: 202–220.
- Smith, S., Avinesh, P. V. S. and A. Kilgarriff. 2010. Gap-fill tests for language learners: corpus-driven item generation. *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*. Held: 8–11 December 2010, Kharagpur, India. Noida: Macmillan Publishers.
- Song, M. Y. 2012. Note-taking quality and performance on an L2 academic listening test. *Language Testing* 29: 67–89.
- Spolsky, B. 1981. Some ethical questions about language testing. In C. Klein-Braley and D. K. Stevenson (Eds.). *Practice and problems in language testing 1*. Frankfurt: Verlag Peter D. Lang.
- Stansfield, C. 2008. Where we have been and where we should go. *Language Testing* 25: 311–326.

- Stansfield, C. W. and W. E. Hewitt. 2005. Examining the predictive validity of a screening test for court interpreters. *Language Testing* 22: 438–462.
- Storey, P. 1997. Examining the test-taking process: a cognitive perspective on the discourse cloze test. *Language Testing* 14: 214–231.
- Sun, Y. and E. Doman. 2018. Peer assessment. In Liantas, J. I. (Ed.). 2018. *The TESOL encyclopedia of English language teaching*. New Jersey: John Wiley and Sons.
- Suzuki, Y. 2015. Self-assessment of Japanese as a second language: the role of experiences in the naturalistic acquisition. *Language Testing* 32: 63–81.
- Tauroza, A. and D. Allison. 1990. Speech rates in British English. *Applied Linguistics* 11: 90–105.
- Taylor, L. 2009. Developing assessment literacy. *Annual Review of Applied Linguistics* 29: 21–36.
- Taylor, L. 2012. Accommodation in language testing. In Coombe, C., Davidson, P., O'Sullivan, B. and S. Stoyloff (Eds.). 2012. *The Cambridge Guide to Second Language Assessment*. Cambridge: Cambridge University Press.
- Trites, L. and M. McGroarty. 2005. Reading to learn and reading to integrate: new tasks for reading comprehension tests? *Language Testing* 22: 174–210.
- Trudgill, P. and J. Hannah. 2017. *International English: a guide to the varieties of standard English (6th edition)*. Oxford: Routledge.
- Tsagari, D. (Ed.). 2016. *Classroom-based Assessment in L2 Contexts*. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Tuzlukova, V. 2018. Projects. In Liantas, J. I. (Ed.). 2018. *The TESOL encyclopedia of English language teaching*. New Jersey: John Wiley and Sons.
- van Ek, J. A. and J. L. M. Trim. 2001a. *Waystage 1991*. Cambridge: Cambridge University Press.
- van Ek, J. A. and J. L. M. Trim. 2001b. *Threshold 1991*. Cambridge: Cambridge University Press.
- van Ek, J. A. and J. L. M. Trim. 2001c. *Vantage*. Cambridge: Cambridge University Press.
- Van Moere, A. 2006. Validity evidence in a university group oral test. *Language Testing* 23: 411–440.
- Wadden, P. and R. Hilke. 1999. Polemic gone astray: a corrective to recent criticism of TOEFL preparation. *TESOL Quarterly* 33: 263–270.
- Wagner, E. 2010. The effect of the use of video texts on ESL listening test-taker performance. *Language Testing* 27: 493–513.

- Wall, D. 1996. Introducing new tests into traditional systems: insights from general education and from innovation theory. *Language Testing* 13: 334–354.
- Wall, D. and J. C. Alderson. 1993. Examining Washback: the Sri Lankan impact study. *Language Testing* 10: 41–69.
- Wall, D. and T. Horák. 2006. *The Impact of Changes in the TOEFL Examination on Teaching and Learning in Central and Eastern Europe: Phase 1, The baseline study*. (Available online at: <https://www.ets.org/Media/Research/pdf/RR-06-18.pdf>)
- Wall, D. and T. Horák. 2008. *The Impact of Changes in the TOEFL Examination on Teaching and Learning in Central and Eastern Europe: Phase 2, Coping With Change*. (Available online at: https://www.ets.org/research/policy_research_reports/publications/report/2008/hspw)
- Wall, D. and T. Horák. 2011. *The Impact of Changes in the TOEFL Examination on Teaching and Learning in Central and Eastern Europe: Phase 3, The Role of the Coursebook. Phase 4, Describing Change*. (Available online at: <https://www.ets.org/Media/Research/pdf/RR-11-41.pdf>)
- Wall, D., Clapham, C. and J. C. Alderson. 1994. Evaluating a placement test. *Language Testing* 11: 321–344.
- Watanabe, Y. 1996. Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing* 13: 318–333.
- Weigle, S. C. 1994. Effects of training on raters of ESL compositions. *Language Testing* 11: 197–223.
- Weigle, S. C. 2002. *Assessing Writing*. Cambridge: Cambridge University Press.
- Weigle, S. C. 2013. English as a second language writing and automated essay evaluation. In *Handbook of automated essay evaluation: Current applications and new directions*. Shermis, M. D. and J. Burstein (Eds.).
- Weir, C. J. 1988. The specification, realization and validation of an English language proficiency test. In Hughes, A. (Ed.). 1988b. *Testing English for university study. ELT Documents 127*. Oxford: Modern English Press.
- Weir, C. J. 1990. *Communicative Language Testing*. Hemel Hempstead: Prentice Hall.
- Weir, C. J. 2005. *Language testing and validation: an evidence-based approach*. Palgrave Macmillan.
- Weir, C. J. and D. Porter. 1995. The Multi-Divisible or Unitary Nature of Reading: the language tester between Scylla and Charybdis. *Reading in a Foreign Language* 10: 1–19.
- Weir, C. J., Hughes, A. and D. Porter. 1993. Reading skills: hierarchies, implicational relationships and identifiability. *Reading in a Second Language* 7: 505–510.

- Weir, C. J., Huizhong, Y. and J. Yan. 2002. *An empirical investigation of the componentiality of L2 reading in English for academic purposes*. Cambridge: Cambridge University Press.
- Wen, Z., Skehan, P., Biedron, A., Li, S. and R. L. Sparks (Eds). 2019. *Language aptitude: advancing theory, testing, research and practice*. New York: Routledge.
- Wigglesworth, G. 1993. Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing* 10: 305–335.
- Woods, A. and R. Baker. 1985. Item response theory. *Language Testing* 2: 119–140.
- Woods, A., Fletcher, P. and A. Hughes. 1986. *Statistics in language studies*. Cambridge: Cambridge University Press.
- Wu, Y. 1998. What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing* 15: 21–44.
- Xi, X. 2010. How do we go about investigating test fairness? *Language Testing* 27: 147–170.
- Yan, X., Maeda, Y., Lv, J. and A. Ginther. 2016. Elicited imitation as a measure of second language proficiency: a narrative review and meta-analysis. *Language Testing* 33: 497–528.
- Youn, S. J. 2015. Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing* 32: 199–225.
- Zhang, Y. and C. Elder. 2011. Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing* 28: 31–50.

Testing for Language Teachers: Author Index

- Abraham, R. G. 203
Adams, M. L. 136
AERA 77
Al-Hamly 7
Al Mahrooqi, R. 231
Alderson, J. C. 27, 28, 38, 39, 61, 76, 161, 162, 190
Allan, A. 162
Allen, H. B. 28
Allison, D. 165
Álvarez, E. 61
Amini, M. 86
Anastasi, A. 38, 56
Anderson, J. C. 101, 103, 104
Anderson, N. J. 232
Andrews, S. 62
Arnold, J. 175
Aryadoust, V. 162
Avinesh, P. V. S. 86
Babaii, E. 232
Bachman, L. F. 38, 39, 76
Bailey, K. M. 61
Baker, R. 251, 255
Barkhuizen, G. 114
Batty, A. O. 174
Beguín, A. 114
Bejarano, Y. 61
Benzehra, R. 7
Bernstein, J. 138
Berwick, R. 39
Biedron, A. 8
Bouwer, R. 114
Boyd, K. 6
Bradshaw, J. 39
Brennan, R. L. 56
Brown, A. 6, 138, 231
Brown, J. D. 27, 56, 61, 203, 255
Brumfit, C. J. 28
Brunfaut, T. 27
Buck, G. 39, 76, 174, 175
Burststein, J. 114
Butler, Y. G. 232
Bygate, M. 116
Cai, H. 174
Cameron, L. 226
Campbell, R. N. 28
Canale, M. 28
Carey, M. D. 138
Carpenter, K. 226
Carroll, J. B. 28
Caudwell, G. 210, 221, 223, 226
Chalhoub-Deville, M. 28
Chapelle, C. A. 28, 113, 190, 203
Cheng, J. 138
Cheng, L. 62, 231
Cheng, W. 232
Chester, M. D. 7
Chiraramanee, T. 86
Choi, H. 162
Choi, I. C. 62
Chung, Y. R. 190
Clapham, C. 6, 27, 38, 39, 76, 226
Clenton, J. 39
Coffman, W. E. 27, 113
Cohen, A. D. 39
Collins Cobuild 77
Conrad, C. A. 194, 203
Coombe, C. 7, 113, 226, 231
Corson, D. 6, 226
Council of Europe 114, 230
Courtney, R. 174
Criper, C. 39
Crystal, D. 175
Cumming, A. 39, 77
Currie, M. 86
Curtis, A. 62, 232
Dávid, G. 190
Davidson, F. 76, 77
Davidson, P. 7, 226, 231
Davies, A. 4, 6, 28, 39
Davy, D. 175
Denman, C. 231
Deville, C. 28
Doman, E. 232
Donitsa-Schmidt, S. 61
Douglas, D. 28
Drackert, A. 204
Dunn, P. K. 138
Ebel, R. L. 27
Eckes, T. 114

- Edenelbos, P. 226
 Elder, C. 27, 39, 114, 138
 Enright, M. K. 113
 Farhady, H. 203
 Feldt, L. S. 56
 Ferman, I. 61
 Ferne, T. 162
 Ferris, D. R. 114
 Field, J. 174
 Fitzpatrick, T. 39
 Fletcher, P. 45, 255
 Fox, J. 27, 39, 114, 231
 Freedle, R. 162, 175
 Frith, J. R. 136
 Fujii, N. 226
 Fulcher, G. 27, 28, 39, 76, 138
 Genesee, F. 231
 Ginther, A. 174, 204
 Gipps, C. 7
 Godshalk, F. L. 27, 113
 Goldstein, Z. 190
 Grant, L. 77
 Graves, B. 27, 114
 Green, A. B. 17, 27, 62, 162
 Gülçur, L. 113
 Gürel, P. 113
 Hale, G. A. 174
 Hamp-Lyons, L. 6, 61, 114
 Hannah, J. 151
 Harding, L. 27, 175
 Harris, D. P. 98, 101
 Harsch, C. 204
 Hartfield, V. F. 103, 105, 113
 Hartig, J. 204
 Hasselgren, A. 207, 226
 Hasselgreen, A. 210, 221, 223, 224
 Heaton, J. B. 86
 Hegelheimer, V. 190
 Hewitt, W. E. 39
 Hilke, R. 61
 Horák, T. 62
 Hubley, N. J. 161
 Hudson, T. 27, 56, 61
 Hughes, A. 7, 27, 28, 33, 39, 45, 61, 113, 161, 162, 175, 179, 203, 255
 Hughes, J.
 Hughey, J. B. 103, 105, 113
 Huhta, A. 7
 Huizhong, Y. 161
 Hyland, F. 114
 Hyland, K. 114
 Ibrahim-González, N. 86
 In'nami, Y. 162
 Inbar, O. 174
 Ioannou-Georgiou, P. 226
 Jacobs, H. L. 103, 105, 113
 Jafarpur, A. 204
 Jamieson, J. M. 113
 Jang, E. E. 27
 Jennings, M. 27, 114
 Johnson, D. M. 114
 Johnson, K. 28
 Kalaja, P. 7
 Kane, M. 38, 39
 Kasper, G. 139
 Kataoka, H. 226
 Katz, A. 7
 Kenyon, D. M. 27
 Keramati, M. N. 203
 Kilgariff, A. 86
 Kim, H. 27, 39
 Klein-Braley, C. 204
 Knoch, U. 27, 114
 Kobayashi, M. 162
 Koizumi, R. 162
 Kokhan, K. 27
 Kostin, I. 162, 175
 Krekeler, C. 162
 Kremmel, B. 39
 Kunnan, A. J. 39
 Lado, R. 18, 43, 56, 204
 Lam, R. 7
 Laufer, B. 190
 Lazaraton, A. 138
 Lee, J. 232
 Lee-Ellis, S. 204
 Leung, S. W. 226
 Lewkowicz, J. A. 27
 Li, S. 8
 Linn, R. L. 38, 56
 Liontas, J. I. 7, 231, 232
 Lloyd, D. 255
 Lumley, T. 138, 162
 Luoma, S. 232
 Luxia, Q. 62
 Lv, J. 204
 Lynch, B. 27
 Maeda, Y. 204
 Mannell, R. H. 138
 Matsumo, S. 232

- McCombie, T. 113
 McGroarty, M. 162
 McKay, P. 226
 McNamara, T. 6, 7, 255
 McNamara, T. F. 138
 Messick, S. 38, 61
 Midraj, J. 232
 Morrow, J. 28
 Morrow, K. 26, 28, 39
 Mulcahy-Ernt, P. 77
 Muñoz, A. P. 61
 Nakatsuhara, F. 138
 Nitko, A. J. 27, 32, 38, 44, 56
 Nixon, C. 224
 North, B. 114
 O'Loughlin, K. 138
 O'Loughlin, K. J. 138
 O'Sullivan, B. 7, 226
 Ockey, G. 175
 Ockey, G. J. 138, 174
 Oller, J. W. 28, 194, 203
 Palmer, A. S. 38, 76
 Pashmforoosh, R. 232
 Pendar, N. 190
 Percicsich, R. 162
 Pilliner, A. 27
 Pitkänen-Huhta, A. 7
 Popham, W. J. 27
 Portal, M. 27, 33, 39, 45, 255
 Porter, D. 28, 33, 39, 162
 Powers, D. E. 77
 Purmensky, K. 231
 Qian, D. D. 190
 Quirke, P. 232
 Raatz, U. 204
 Raven, J. 7
 Read, J. 161, 190
 Rea-Dickens, P. 6, 226
 Reves, T. 61
 Riazi, M. 39
 Rimmer, W. 190
 Rixon, S. 206, 224, 226
 Roeever, C. 6, 139
 Rupp, A. A. 162
 Ryan, K. 7
 Sadhwani, P. 232
 Saif, S. 62
 Saito, H. 232
 Sanders, T. 114
 Schedl, M. 190
 Schneider, G. 114
 Scott, M. L. 27
 Scott-Barrett, F. 161
 Shaw, S. D. 113
 Sheetz, D. 232
 Shehadeh, A. 232
 Sherman, J. 174
 Shermis, M. D. 114
 Shizuka, T. 162
 Shohamy, E. 6, 7, 27, 61, 114, 174
 Skehan, P. 8, 27
 Smith, S. 86
 Song, M. Y. 174
 Sparks, R. L. 8
 Spolsky, B. 6
 Stansfield, C. 7
 Stansfield, C. W. 27, 39
 Stevenson, D. K. 6
 Storey, P. 39, 203
 Stoyhoff, S. 7, 226
 Sun, Y. 232
 Suzuki, Y. 232
 Swain, M. 28
 Swineford, F. 27, 113
 Szabo, G. 162
 Taghaddomi, S. 232
 Takeuchi, O. 162
 Tarnanen, M. 232
 Tatsuoaka, K. 174
 Taurosa, S. 165
 Taylor, L. 7, 39
 Teasdale, A. 226
 Timukova, A. 204
 Tomlinson, M. 224
 Trim, J. L. M. 177, 190
 Trites, L. 162
 Trudgill, P. 151
 Tsagari, D. 231
 Tuzlukova, V. 232
 Ünal, A. 162
 Upshur, J. 231
 Urbina, S. 38, 56
 Van den Bergh, H. 114
 van Ek, J. A. 177, 190
 Van Moere, A. 138
 Vinjé, M. P. 226
 von Randow, J. 114
 Wadden, P. 61
 Wagner, E. 174, 175
 Wall, D. 27, 38, 39, 61, 62, 76

- Warren, M. 232
Watanabe, J. 62
Watanabe, Y. 61
Weigle, S. C. 113, 114
Weir, C. 162
Weir, C. J. 27, 28, 33, 38, 39, 76,
113, 161, 162
Wen, Z. 8
Wigglesworth, G. 138
Woods, A. 45, 251, 255
Wormuth, D. R. 103, 105, 113
Wu, Y. 39
Xi, X. 39
Xu, J. 190
Yan, J. 161
Yan, X. 204
Yashima, T. 162
Yoshizawa, K. 162
Youn, S. J. 139
Yu, Y. 62
Zhang, L. 162
Zhang, Y. 138
Zingraf, S. A. 103, 105, 113

Testing for Language Teachers: Subject Index

- Academic Word List 85
- accent
 - listening tests 166, 175
 - making tests reliable 51
 - specifying test content 66
 - test scales 134, 135
- accuracy in measurement 1, 2
- achievement tests 12–15
- ACTFL (American Council for the Teaching of Foreign Languages) 28, 145
- addressees 65, 72, 88, 89
- administration of tests
 - preparation 238–241
 - reliability 40, 43
 - uniform and non-distracting conditions 52
- ALTE (Association of Language Testers in Europe) 28
- alternate form, estimates of reliability 44, 247
- ambiguous instructions 52
- ambiguous items 51
- anagrams 219
- analytic scoring 101–105, 134
- anchor items 69, 254
- ANOVA (Analysis of Variance) 246
- aptitude in language 8
- articles (e.g. *the*, *a*) 16, 179
- artificial intelligence 234
- assessment literacy 5, 7
- authenticity 27, 97, 144, 166, 235
- automated scoring 107–108, 110, 114, 234–235
- automated speaking tests 138

- backwash 3–4, 57–62, 142, 209
- benchmark scripts 108
- BNC (British National Corpus) 77, 178, 190

- C-test 199–200
- calibration 109
- CAEL (Canadian Academic English Language) Assessment 27
- Cambridge English B2 First 87, 92, 110, 115, 118, 132, 181
- Cambridge English C2 Proficiency 192, 196, 203
- Cambridge Grammar Profile 177
- Cambridge English Vocabulary Profile 185
- Cambridge English 'Write and Improve' 230
- Cambridge IELTS 21–23
- Cambridge Young Learners English Test 216, 219
- CEFR *see* Council of Europe
- change of topic, speaking ability 125
- channel 66
- cheating 240
- classical item analysis 248–250
- classroom assessment 228, 231
- cloze
 - C-test 199–200
 - conversational 197–199
 - dictation 200–201
 - further reading 203
 - multiple choice 196–197
 - selected deletion 194–196
 - summary 151
 - traditional 193–194
- COBUILD 77
- communicative competence 139
- communicative language testing 26, 28
- comparative judgement 54, 109
- compensation, completion items 101
- computer adaptive testing 25, 46, 234
- computer-based tests 24–25, 28, 235–236
- content of test 1, 13, 29–30, 65–66
- conversational cloze 197–199
- correlation coefficient 31, 32, 249
- corpora 85
- Council of Europe 114, 230
- criterial levels of performance 67, 73, 145, 166–174

- criterion referenced testing 20–24, 27, 145, 254–255
- C-Tests 199–200
- decision consistency 47, 56
- definition items 186, 188
- diagnostic tests 15–16, 27, 141, 177
- DIALANG 16, 26, 27, 190, 230
- dictation
- partial 172–173, 201
 - as testing technique 200–201, 204
- direct testing 17–19, 27, 36, 58–59
- discontinuous grammatical elements 84, 183
- discrete point testing 19–20, 28
- discrimination indices 248–250, 254
- discrimination, lack of 249–250
- discussion in oral testing 127
- distractors 80–82, 148, 186, 250
- EAQUALS 177
- elicited imitation 201–202, 204
- error analysis 104–105
- error gravity 105
- ETS (Educational Testing Services) 39, 114
- ethics 6
- face-to-face tests 25
- facility values 248, 250, 254
- fairness 37, 39
- FBI listening summary translation exam 27
- FCE *see* Cambridge English B2 First
- feedback to candidates 109–113, 207, 227, 229
- fit of items and people 251
- Flesch-Kincaid Grade Level Score 66
- Flesch Reading Ease Score 66
- flexibility 25, 67, 124
- formative assessment 15, 82, 110, 227
- format of test 24–28, 52
- see also* techniques
- freedom, restricting candidates 50–51, 53
- frequency table 242–243
- FSI (Foreign Service Institute) 134
- gap filling items 84–86, 150–156, 170, 178–179, 188, 220–223
- GCSEs (General Certificate of Secondary Education) 59
- general knowledge 92, 146, 157
- grammar testing
- completion technique 182–183
 - multiple choice technique 183
 - paraphrasing technique 180–182
 - reasons for testing grammar separately 176–177
 - sampling 178
 - scoring 184
 - writing items 178–179
 - writing specifications 177
- graphic features of texts 144
- group oral 138
- guessing 2, 79–80
- halo effect 103
- handbooks, test 28, 70–71, 77, 259
- histogram 243–244
- holistic scoring 98, 105, 109, 134
- identification of candidates 54
- ILR (Interagency Language Roundtable) 20, 101, 132, 145
- impact in educational measurement 3
- independence of items 157
- indirect testing 17–19, 34–35, 36
- inferences 19, 143
- information requests 125
- information transfer 154–156, 170–171, 215–216
- innovation theory 61
- instructions 51–52, 93, 240–241
- integrative testing 19–20, 28
- interaction with fellow candidates 117, 234
- internal consistency, co-efficient of 44
- interpreting tasks 126
- interruption, candidate responses to 125
- interview 41–42, 48, 124–125
- introspection 39, 175
- see also* retrospection
- invitation to ask questions 125
- item analysis 247–254

- item banking 81, 236, 256–257
- item response theory (IRT) 46, 247–248, 250–251
- item–test correlation 31, 200–201, 203, 249
- items, number of 66
- language assessment literacy 5, 7
- language of items and responses 67, 68
- layout *see* format of test
- listening, testing of
 - critical levels 166
 - scoring 174–175
 - as separate skill 163–166
 - specifying what the candidate should be able to do 163–166
 - task setting 166–167
 - techniques 168–174
 - writing items 167–168
 - young learners 211–216
- mean 244–245
- median 244–245
- medium 66, 75
 - see also* format of test
- Michigan Test 28
- modal verbs 16, 84
- mode 244–245
- moderation
 - checklist 156
 - listening tests 171, 173
 - test development 68, 69, 73
 - trialling 69
- monologue 126
- multiple choice cloze 196–197
- multiple choice questions 86, 147–148, 162, 168–169, 183, 216–217
- needs analysis 89
- normal distribution 45
- norm referenced testing 20–21, 24, 27, 145, 254
- Norway 206
- notes as basis of writing task 89, 90
- note taking 171–172, 174
- objectives of the course 31, 59–60
- objective testing 24, 54
- open-ended items 53
- operations, demonstrated by
 - candidates 65, 163–164, 176, 185
- options, multiple choice 78, 80, 82, 186
- oral ability *see* speaking tests
- overall ability
 - cloze procedures 193–199
 - as concept 192
 - C-tests 199–200
 - Dictation 200–202
 - measuring 192–193, 202–205
- Oxford 3000 77, 190
- paper-and-pencil tests 25
- paraphrase items 180–182
- passages, number of 49, 66
- past continuous 84, 181
- Pearson 'English Benchmark' 230
- Pearson Global Scale of English 185, 190
- Pearson Test of English (PTE) 39, 201, 203
- Pearson Test of English for Young Learners 94, 224
- peer-assessment 230
- pictures, use of 94–95, 125, 187, 209, 211, 219–224
- placement tests
 - online resources 28
 - overall ability tests 202
 - predictive validity 33
 - reading tests 141
 - statistical analysis 242–243, 248
 - test development 74
 - use of 16–17, 27
 - vocabulary tests 184, 185
 - writing tests 177
- pop quizzes 15
- portfolio assessment 113
- practice materials 259
- predictive validity 32–33
- proficiency tests
 - distinction from achievement tests 14
 - grammar tests 176
 - online resources 28
 - online security 26
 - predictive validity 32–33
 - reading tests 141
 - screening 17
 - test development 248

- use of 11–12
- vocabulary tests 184, 185, 187
- questions, choice of 50
 - see also* multiple choice
 - questions; short-answer
 - questions
- range of scores 105–106, 244
- Rasch analysis 251–254
- rating *see* scales; scoring
- readership, intended 144
- reading aloud 126–127
- reading, testing
 - expeditious 72–73, 141–142, 145
 - scoring 145, 158–161
 - setting the tasks 146–147
 - slow and careful 72–73, 141–143, 145
 - specifying what the candidate should be able to do 140–145
 - speed of 66, 145
 - writing items 147–157
 - young learners 216–218
- recordings as test stimuli 127, 166–167, 175
- reduced redundancy 192–193
- referents, identifying 148
- reliability 40–42
 - coefficient 42–45, 48, 245–246, 250
 - computer use 234
 - estimates of 43–44, 46–47, 56, 247
 - improving 49–55
 - inter-scorer 48, 109
 - intra-scorer 48
 - lack of 2–3
 - overall ability tests 202
 - scorers 47–48
 - standard error of measurement and the true score 44–47
 - statistical analysis 245–247, 250
 - and validity 38, 55
- requests for information and elaboration 125
- residuals 252
- retrospection 35
 - see also* introspection
- role play 125–126, 127
- sample tests 52, 60
- sampling behaviour 19, 49–50, 78
- sampling specifications 58
- scales
 - calibration 70
 - reading 145
 - self assessment 230
 - speaking 132, 134, 136
 - writing 98, 99, 101, 104–106, 113, 114
 - validity 34
- scanning skills 146, 148–149, 151, 157
- scoring
 - calibration 70, 106, 136
 - grammar 184
 - listening tests 174–175
 - peer assessment 230
 - reading tests 145, 158–161
 - self-assessment 230
 - speaking tests 129–137
 - writing tests 97–109
- screening tests 17, 202
- search reading 142, 146
- second language acquisition 39, 190
- selected deletion cloze 194–196
- self-assessment 207, 230, 232
- semi-direct testing 19, 127, 138
- short-answer items 83
- sequencing items 104, 150, 158
- simulated conversation 128
- size of response, discrimination indices 249–250
- skills
 - informational 116, 164
 - interactional 116–117, 164–165
 - in managing interactions 117
 - sub- 33–34, 65, 158, 162
- Socratic 230
- sub-skills 33–34, 65, 158, 162
- skimming 30, 142
- speaking, testing of
 - challenges 115
 - representative tasks 115–124
 - scoring 129–137
 - valid sampling 124–128
 - young learners 223–225
- Spearman–Brown formula 44, 49, 246

- specifications
 - criterion referenced testing
 - 20–24, 27, 145, 254–255
 - listening 163–166
 - reading 140–145
 - sampling 58
 - test development 64–67
 - validity 30–33
- speech, rate of 117, 134, 135, 165
- speed of computer processing 233
- speed of processing for reading 66, 145
- split half estimate of reliability 44
- stages of test development *see* test development
- stakeholders 5
- standard deviation 245, 246
- standard error of individual's score 45, 46
- standard error of measurement (SEM) 44–46, 247, 253
- statistical analysis
 - criterion-referenced tests 254–255
 - item analysis 247–254
 - reliability 245–246
 - scores 242–245
 - use of 242
- stems (test questions) 78, 186
- structural range 66, 75
- structure of tests *see* test structure
- subjective testing 24, 48, 60
- summative assessment 15, 110
- syllabus content approach 13
- synonym items 185–186
- techniques, test 66
- TEDDCLOG 86, 191
- TEEP (Test of English for Educational Purposes) 76
- test development
 - handbook writing 70–71
 - informal trialling 69
 - maintenance of test 71–76
 - procedures 63–64
 - rating scales 70
 - stating the problem 64
 - trailing 69–70
 - training 71
 - validation 70
- writing and moderating
 - items 67–68
 - writing specifications 64–67
- test features 209–210
- test format 24–28, 52
- test purposes 4–5, 8–9
- test-retest estimate of reliability 43
- test security 26, 69, 236
- test structure 66
- test techniques 66
- test wiseness 162
- text form 65–66
- text length 66
- text types 65
- texts, selection of 13, 66
- think-aloud 39
- timing 52, 66, 73
- TOEFL (Test of English as a Foreign Language) 6, 39, 61–62, 99, 118, 123
- topics 66
- traditional cloze 193–194
- training of staff 106–107, 114, 131–132, 137
- transcription 173, 216
- trailing 69–70, 73–74, 254, 256–257
- True/False* items 82, 148, 161
- true score 44–47, 247, 254
- TWE (Test of Written English) 6, 98
- unfamiliar words, predicting
 - meaning of 72, 185
- unique answer items 150, 170
- validation 30–31, 36–37, 39, 70, 203
- validity 29
 - coefficient 31–32, 33
 - concurrent 31–32, 34
 - consequential 37–38
 - construct 33–35
 - content 29–30, 34
 - criterion-related 30
 - face 36
 - fairness 37
 - making tests more valid 36–37
 - predictive 32–33
 - and reliability 38, 55
 - in scoring 35–36; *see also* scoring
 - speaking tests 124–128

- writing tests 92–97
- Versant English Test 129
- vocabulary range 66, 144
 - production ability 187–188
 - sampling 185
 - writing tests 185–187
 - testing of 135, 151, 184–190
- washback as term 57
 - see also* backwash
- weighting 103, 105, 134, 136
- writing, testing of
 - comparative judgement 109
 - feedback 109–113
 - following acceptable procedures 108–109
 - representative tasks 87–92
 - scoring 97–109
 - valid sampling of ability 92–97
 - young learners 219–223
- Yes/No items 82, 125
- young learners tests
 - reasons for and general approach 206–208
 - recommended techniques 210–226
 - specific features 209–210