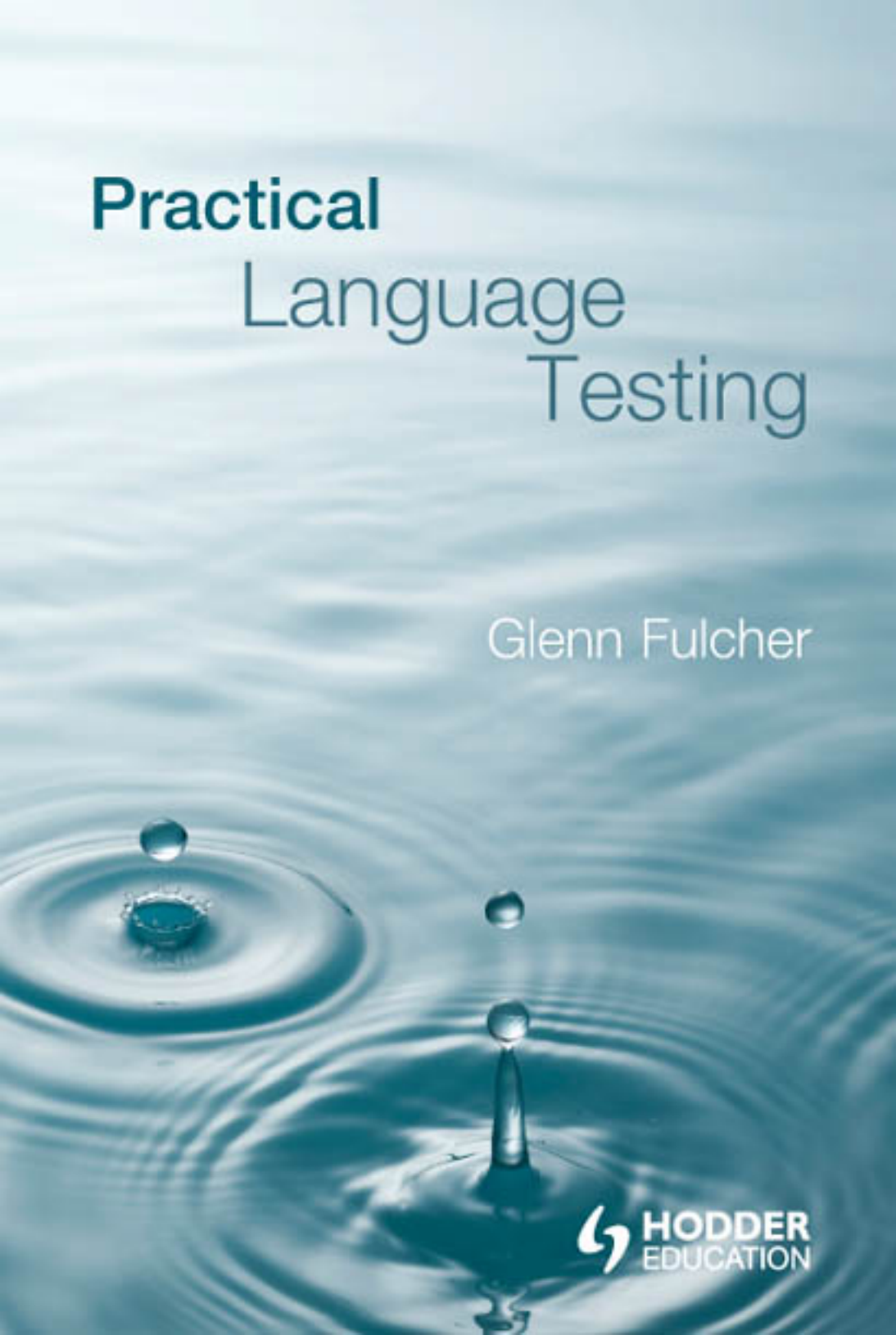
Practical Language Testing

Glenn Fulcher



HODDER EDUCATION

Practical Language Testing

Glenn Fulcher

HODDER EDUCATION
AN HACHETTE UK COMPANY

For all the inspiring teachers
I have been lucky enough to have
and especially
Revd Ian Robins
Who knows where the ripples end?

First published in Great Britain in 2010 by
Hodder Education, An Hachette UK Company,
338 Euston Road, London NW1 3BH

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978 0 340 984482

1 2 3 4 5 6 7 8 9 10

Cover Image © Anthony Bradshaw/Photographer's Choice RF/Getty Images
Typeset in 10 on 13pt Minion by Phoenix Photosetting, Chatham, Kent
Printed and bound in Great Britain by Antony Rowe, Chippenham, Wilts

What do you think about this book? Or any other Hodder Education title? Please send your comments to educationenquiries@hodder.co.uk

http://www.hoddereducation.com

Contents

Acknowledgements

I am deeply indebted to the Leverhulme Trust (www.leverhulme.ac.uk), which awarded me a Research Fellowship in 2009 in order to carry out the research required for this book, and funded study leave to write it. The generosity of the Trust provided the time and space for clear thinking that work on a text like this requires.

The University of Leicester was extremely supportive of this project, granting me six months' study leave to work entirely on the book. I would also like to thank staff in the School of Education for help and advice received while drafting proposals and work schedules.

I am grateful to the people, and the institutions, who have given me permission to use materials for the book.

Special thanks are due to Professor Yin Jan of Shanghai Jiao Tong University, and Chair of the National College English Testing Committee of the China Higher Education Department. Her kindness in providing information about language testing in China, as well as samples of released tests, has enriched this book.

I have always been inspired by my students. While I was working on the development of Performance Decision Trees (see Chapter 7), Samantha Mills was working on a dissertation in which she developed and prototyped a task for use in assessing service encounter communication in the tourist industry. In this book the two come together to illustrate how specifications, tasks and scoring systems, can be designed for specific purpose assessment. I am very grateful to Samantha for permission to reproduce sections of her work, particularly in Chapters 5 and 6.

Test design workshops can be great fun; and they are essential when brainstorming new item types. I have run many workshops of this kind, and the material used to illustrate the process of item evaluation in Chapter 6 is taken from a workshop conducted for Oxford University Press (OUP). I am grateful to OUP, particularly Simon Beeston and Alexandra Miller, for permission to use what is normally considered to be confidential data.

The book presents a number of statistical tools that the reader can use when designing or evaluating tests. All of the statistics can be calculated using packages such as SPSS, or online web-based calculators. However, I believe that it is important for people who are involved in language testing to understand how the basic statistics can be calculated by hand. My own initial statistical training was provided by Charles Owen at the University of Birmingham, and I have always been grateful that he made us do calculations by hand so that we could 'see' what the machine was doing. However, calculation by hand can always lead to errors. After a while, the examples in the text became so familiar that I would not have been able to spot any errors, no matter how glaring. I am therefore extremely grateful to Sun Joo Chung of the University of Illinois at Urbana-Champaign for the care with which she checked and corrected these parts of the book.

The content of the book evolved over the period during which it was written. This is because it is based on a research project to discover the language testing needs of teachers and students of language testing on applied linguistics programmes. A survey instrument was designed and piloted, and then used in the main study. It was delivered through the Language Testing Resources website (http://languagetesting.info), and announced on the language testing and applied linguistics discussion lists. It was also supported by the United Kingdom's Subject Centre for Languages, Linguistics and Area Studies. The respondents came from all over the world, and from many different backgrounds. Each had a particular need, but common themes emerged in what they wished to see in a book on practical language testing. The information and advice that they provided has shaped the text in many ways, as my writing responded to incoming data. My thanks, therefore, to all the people who visited my website and spent time completing the survey.

My thanks are also due to Fred Davidson, for a continued conversation on language testing that never fails to inspire. To Alan Davies and Bernard Spolsky, for their help and support; and for the constant reminder that historical context is more important than ever to understanding the 'big picture'. And to all my other friends and colleagues in the International Language Testing Association (ILTA), who are dedicated to improving language testing practice, and language testing literacy.

Every effort has been made to obtain the necessary permission with reference to copyright material. The publishers apologise if inadvertently any sources remain unacknowledged and will be glad to make the necessary arrangements at the earliest opportunity.

Finally, acknowledgements are never complete with recognition for people who have to suffer the inevitable lack of attention that writing a book generates. Not to mention the narrowing of conversational topics. My enduring thanks to Jenny and Greg for their tolerance and encouragement.

Figures

This page intentionally left blank

Tables

This page intentionally left blank

Preface

This book is about building and using language tests and assessments. It does what it says on the tin: it is a *practical* approach. However, it does not provide ready-made solutions. Language testing is a complex social phenomenon, and its practice changes lives. The book therefore assumes that you will wish to think carefully about testing and its impact in your own context.

The term 'practical' therefore needs some definition. The book is 'practical' in the sense that it gives guidance on how to do things to build a test. It is also 'practical' in that each chapter will be useful to you when you come to making decisions about when, why and how to conduct assessments. The book is designed to provide the *knowledge* you will need to apply, and the *skills* you will need to practise. However, if we are to build good language tests, we have to be aware of the larger social, ethical, and historical context, within which we work. If language testing and assessment are not guided by *principles*, we could end up doing more harm than good. Davies (2008a) has cogently argued that testing and assessment texts that do not embed knowledge and skills in principles ignore the increasing demand of professionalism and social responsibility.

Language professionals, applied linguists and educational policy makers need an expanded 'assessment literacy' in order to make the right decisions for language learners and institutions (Taylor, 2009). This literacy will be about learning the nuts and bolts of writing better test items (Coniam, 2008), and establishing a core knowledge base (Inbar, 2008); but it is also about appreciating the reasons why we test, why we test the way we do and how test use can enrich or destroy people's hopes, ambitions and lives.

Although I am far from being in the 'postmodern' school of language testing and assessment, the view that language testing is a social activity cannot be denied (McNamara, 2001). Nor can the fact that our practices are thoroughly grounded in a long history that has brought us to where we are (Spolsky, 1995). It is partly because of this history that many texts published 'for teachers' focus almost entirely upon the technologies of normative large-scale standardised testing. While it is important that teachers are familiar with these, they are not always directly relevant to the classroom. This book therefore tries to introduce a balance between standardised testing and classroom assessment.

The structure reflects a conscious decision to place language testing and assessment within context, *and* to provide the 'practical' guidance on the nuts and bolts of test building. Broadly, the first three chapters survey the language testing landscape upon which we can build. Chapter 4 is about the material that we can use in construction, and the rest of the book takes the reader through the process of building and implementing a language test.

Chapter 1 considers the purpose of testing in the broadest sense of why societies use tests, and in the narrow sense of how we define the purpose of a particular test. It looks

at how tests are used, for good and ill; and the unintended consequences that testing can have on people who are caught up in the need to succeed. Chapters 2 and 3 deal in turn with large-scale standardised testing, and then with classroom assessment. The stories of both paradigms are set within a historical framework so that you can see where the theories and practices originate.

In Chapter 4 we begin the journey through the process of test design, starting with deciding what to test, and why. Chapter 5 begins the test design process in earnest, where we discuss how to create test specifications – the basic design documents that help us to build a test. This is where we learn to become 'test architects', shaping the materials and putting them together in plans that can be used to produce usable test forms. In Chapter 6 we look at how to evaluate the test specifications and test items, from initial critical discussions in specification workshops to trying out items and tests with learners. Chapter 7 contains a discussion of scoring, covering both traditional item types like multiple choice, as well as performance tests that require human judgement. Frequently, we have to use tests to make decisions that require a 'cut score' – a level on the test above which a test taker is judged to be a 'master', and below which they are still 'novices'. Establishing cut scores and linking these to absolute standards is the subject of Chapter 8. Chapter 9 discusses the practicalities of test administration, and why the 'rituals' of testing have grown as they have.

Finally, in Chapter 10, we return to the classroom and to the effect that tests have upon learning and teaching, and how we go about preparing learners to take tests.

Throughout the book I have included examples from real tests and assessments. Some of these are good examples that we can emulate. Others are provided for you to critique and improve. Some of them are also drawn from historical sources, as 'distance' is useful for nurturing critical awareness. However, I do not present sets of typical test items and tasks that you could simply select to include in your own tests. There are plenty of books on the market that do this. This book asks you to think about what item or task types would be most useful for your own tests. We discuss options, but only you can provide the answers and the rationales for the choices you make.

There are activities at the end of every chapter that you can attempt on your own, although many would benefit from team work. Sharing experiences and debating difficult issues is best done in a group. And it's also more fun. The activities have been designed to help you think through issues raised in the chapter, and practise the skills that you have learned. The activities are not exhaustive, and you are encouraged to add to these if you are using the text in a language testing course. Beginning in Chapter 4 there is also a Project that you may wish to do as you move from chapter to chapter.

This structure has been shaped not only by my own understanding of what an introductory book to foster 'assessment literacy' might look like, but also by what language teachers and students of applied linguistics have told me that they need to know, and be able to do. Prior to writing the book I undertook a large-scale internet-based survey, funded by the Leverhulme Trust. Almost 300 respondents completed the survey, and I was struck by the sophistication of their awareness of assessment issues.

Here is a selection of typical responses to a question about what teachers and students of applied linguistics most need in a 'practical' language text:

Evaluating reliability for our in-house tests, and checking questions at each stage in test development.

I don't understand statistics, but I know they can be useful. I need it explaining conceptually.

We need to know the jargon, but introduce it step by step.

Hands-on activities; examples of test specs; a glossary would be useful.

A book of this type must focus on the basics of item writing and test construction, the basic concepts of validity and reliability, particularly in regards to the assessment of speaking and writing. It must also cover the ethics of test use and test score interpretation.

Developing classroom tests, performance tests, setting score standards, deciding what to test, preparing learners for test situations.

Differentiation between classroom assessments, formative assessment, and large-scale assessment when discussing key issues.

Most of the assessment/testing practices are done by teachers; I think that a book should be aimed at 'normal' language teachers more than specialists in testing, they already have other sources of information and training.

Issues to do with ensuring validity and reliability in language testing. The test writing process from the creation of test specifications through to the trialling, administration and marking of tests.

Vignettes; glossary; application activities for individuals and groups, including some practice with basic test statistics and approaches to calculating grades.

Some information on testing as an industry, a multi-billion dollar concern and why we have to fight crap when we see it.

Luckily, many respondents said they realised that it is impossible to include everything in a practical language testing book. This is evidently true, as you will see. I am sure to have left out a topic that you think should have been included. One respondent understood this all too well: 'The book should be well-structured, clearly focused, and however tempted you might be to put everything into one book, you should be selective in order to be comprehensible and user-friendly.' I am not entirely sure that I have achieved this. But if I have got even halfway there, my time will have been well spent.

As another respondent said, 'The learning never ends.' In order to sustain you during your journey through the book, you may wish to pay regular visits to my website:

http://languagetesting.info

Here you will find a set of online videos that define and explain some of the key concepts and topics in language testing. To help you with additional reading, I have links to online articles, and other language testing websites. There are links to useful journals, and regular updates on testing stories that get into the news.

Constructive criticism is always welcome, via the website.

1 Testing and assessment in context

1. Test purpose

Language testing, like all educational assessment, is a complex social phenomenon. It has evolved to fulfil a number of functions in the classroom, and society at large. Today the use of language testing is endemic in contexts as diverse as education, employment, international mobility, language planning and economic policy making. Such widespread use makes language testing controversial. For some, language tests are *gatekeeping* tools that further the agendas of the powerful. For others, they are the vehicle by which society can implement equality of opportunity or learner empowerment. How we perceive language tests depends partly upon our own experiences. Perhaps they were troubling events that we had to endure; or maybe they opened doors to a new and better life. But our considered judgements should also be based upon an understanding of the historical evolution of testing and assessment, and an analysis of the legitimate roles for testing in egalitarian societies. This first chapter therefore situates language testing in its historical and social context by discussing a variety of perspectives from which to evaluate its practical applications, beginning with the most fundamental concern of all: the purpose of testing.

The act of giving a test always has a purpose. In one of the founding documents of modern language testing, Carroll (1961: 314) states: 'The purpose of language testing is always to render information to aid in making intelligent decisions about possible courses of action.' But these decisions are diverse, and need to be made very specific for each intended use of a test. Davidson and Lynch (2002: 76–78) use the term 'mandate' to describe where test purpose comes from, and suggest that mandates can be seen as either internal or external to the institution in which we work. An internal mandate for test use is frequently established by teachers themselves, or by the school administration. The purpose of such testing is primarily related to the needs of the teachers and learners working within a particular context. Tests that are under local control are mostly used to place learners into classes, to discover how much they have achieved, or to diagnose difficulties that individual learners may have. Although it is very rarely discussed, teachers also use tests to motivate learners to study. If students know they are going to face a quiz at the end of the week, or an end of semester achievement test, the effect is often an increase in study time near the time of the test. In a sense, no 'decision' is going to be taken once the test is scored. Indeed, when classroom tests were first introduced into schools, an increase in motivation was thought to be one of their major benefits. For example, writing in the nineteenth century, Latham (1877:

146) reported: 'The efficacy of examinations as a means of calling out the interest of a pupil and directing it into the desired channels was soon recognized by teachers.' Ruch (1924: 3) was a little more forthright: 'Educators seem to be agreed that pupils tend to accomplish more when confronted with the realization that a day of reckoning is surely at hand.' However, the evidence to support the motivational role of tests has always been largely anecdotal, making it a folk belief, no matter how prevalent it has always been.

The key feature of testing within a local mandate is that the testing should be 'ecologically sensitive', serving the local needs of teachers and learners. What this means in practice is that the outcomes of testing – whether these are traditional 'scores' or more complex profiles of performance – are interpreted in relation to a specific learning environment. Similarly, if any organisational or instructional decisions are taken on the basis of testing, their effect is only local.

Cronbach (1984: 122) put this most succinctly:

> *A test is selected for a particular situation and purpose. What tests are pertinent for a psychological examination of a child entering first grade? That depends on what alternative instructional plans the school is prepared to follow. What test of skill in English usage is suitable for surveying a high school class? Those teachers for whom clarity of expression is important will be discontented with a test requiring only that the student choose between grammatically correct and incorrect expressions.*

If testing with a local mandate is ecologically sensitive, it is highly likely that it will have a number of other distinguishing characteristics. Firstly, we would expect much of the testing to be *formative*. That is, the act of testing is designed to play a role in the teaching and learning process, rather than to certify ultimate achievement. Secondly, the test is likely to be *low-stakes*. This means that any decisions made after the testing is complete will not have serious consequences for the person who has taken the test, for the teacher or for the school. Rather, the information from the testing or assessment procedure will be used by the teacher and the learner to make decisions about what the most immediate learning goals might be, what targets to set for the next semester, or which classes it is most useful for a learner to attend. If mistakes are made, they are easily corrected through dialogue and negotiation. Thirdly, the testing or assessment procedures used are likely to be created or selected by the teachers themselves, and the learners may also be given a say in how they prefer to be assessed. This ecological sensitivity therefore impacts upon how testing is used, the seriousness (and retractability) of decisions, and the involvement of the local *stakeholders* in designing and implementing tests and assessments.

An external mandate, on the other hand, is a reason for testing that comes from outside the local context. The decision to test is taken by a person or a group of people who often do not know a great deal about the local learning ecology, and probably don't even know the teachers and learners who will have to cope with the required testing regime. As soon as we begin to talk about external mandates loaded words begin to enter the discussion, such as 'regime', because teachers are naturally suspicious of

anything that is 'imposed' from outside. The motivations for external mandates may also appear extremely vague and complex; indeed, policy makers often do not clearly articulate the purpose of the required testing, but it usually serves a very different function from internally mandated tests. External tests are primarily designed to measure the proficiency of learners without reference to the context in which they are learning. Also, the tests are *summative*: they measure proficiency at the end of a period of study, by which time learners may be expected to have reached a particular *standard*. The information therefore doesn't always feed back into the learning process, but fulfils an accountability role.

In summative testing we also expect test scores to carry *generalisable* meaning; that is, the score can be interpreted to mean something beyond the context in which the learner is tested. In order to understand this, we can turn to Messick (1989: 14–15), who said that generalisability is about 'the fundamental question of whether the meaning of a measure is context-specific or whether it generalizes across contexts'. Teachers wish the meaning of testing and assessment to be locally meaningful in terms of what comes next in teaching. If the outcomes are not particularly generalisable across people, settings and tasks – or different 'ecological conditions' – it doesn't matter too much. In externally mandated tests, however, there is an assumption that the meaning of test scores generalise to what learners are capable of doing across a wide range of contexts not necessarily contained in the test. Score users want to be able to make decisions about whether learners can communicate with people outside their immediate environment, in unfamiliar places, engaging in tasks that have not been directly modelled in the test itself. The greater the claim for generalisability, the more 'global' the intention to interpret score meaning. For example, an academic writing task may contain only one or two questions, but the scores are treated as being indicative of ability to write in a wide range of genres, across a number of disciplines. Or we could think of scores on a short reading test being used to compare literacy rates across a number of countries. The testers might wish to draw conclusions about the likely contribution of the educational sector to the economy. Indeed, the latter is the explicit aim of the Programme for International Student Assessment (PISA), carried out by the Organisation for Economic Co-operation and Development (www.pisa.oecd.org).

Generalisability is therefore an important consideration in tests with an external mandate, when they are used to certify an ability to perform at a specified level, or to compare and contrast the performance of schools, educational districts, or even countries. We refer to such tests as being *high-stakes*. Failure for individual learners may result in the termination of their studies. Or they may not be able to access certain occupations. For schools, a 'failure' may result in a Ministry of Education introducing 'special measures', including removal of staff, or direct management from the central authority. At the national level, perceived failure in comparison with other countries could result in the wholesale reform of educational systems as politicians try to avoid the implied impending economic catastrophe.

2. Tests in educational systems

One of the largest testing systems in the world is the National College Entrance Test in China (the Gaokao). Taken over a two-day period, students sit tests in Chinese, English, mathematics, sciences and humanities. The outcome is a score that can range between 100 and 900 points, and determines which college or university each student will attend. Each college and university sets its entrance score and allocates a number of places to each province. Millions of students apply for a place, and so the test is extremely high-stakes and very competitive.

Why do such tests exist? Testing is primarily about establishing *ways of making decisions* that are (hopefully) not random, and seen as 'fair' by the population. Whenever we establish ways of making decisions, we reveal what we believe about society and political organisation. So the practice of testing and assessment can never be separated from social and political values.

This may sound like an overstatement. But consider the university application situation again. There are a limited number of places in institutions of higher education and there must be some method of judging which applicants to accept. We could make the acceptance decisions using many different criteria. If the criteria that we use reflect our views about how society is (or should) be organised, what would it say about us if we decided to offer the best places to the children of government officials? Or to those who can pay the highest fees? If you find these two suggestions rather distasteful, perhaps you should ask this question of yourself: what do you think the goals of education are?

Here is another strong statement: 'the act of testing is the mechanism by which our social and political values are realised and implemented.' If we believe that the purpose of a test like the Gaokao is to provide equality of opportunity, we see meritocratic practices embedded within the testing process. Messick (1989: 86–87) was one writer who believed that this was the primary social purpose of testing. He argued that testing, when done well, was capable of delivering 'distributive justice' (Rawls, 1973):

> *If desirable educational programs or jobs are conceived as allocable resources or social goods, then selection and classification may be viewed as problems of distributive justice. The concept of distributive justice deals with the appropriateness of access to the conditions and goods that affect individual well-being, which is broadly conceived to include psychological, physiological, economic and social aspects. Any sense of injustice with respect to the allocation of resources or goods is usually directed at the rules of distribution, whereas the actual source of discontent may also (or instead) derive from the social values underlying the rules, from the ways in which the rules are implemented, or from the nature of the decision-making process itself.*

In the Gaokao there is an assumption that access to university places should be based on a principle of meritocracy that places a high value on ability, as defined by the tests. There is also a clear commitment to equality of opportunity. This means that there should be no discrimination or *bias* against any test taker or group of test takers. We

could question these values, of course. Access to higher education has in the past been a matter of ability to pay, which in many countries was related to class; but social immobility is not something that we would wish to defend today. Other options might be to value effort above ability. Perhaps it is those individuals who strive hard to improve who should be given the better education? We might assess for progress from a baseline, therefore valuing commitment, dedication and staying power. In a world of global business where the principles of capitalism do not seem to be frequently challenged, perhaps the process should merely be opened up to market forces?

What we choose to endow with high value tells us a great deal about what we expect the effects of testing to be. It has even been argued that *effect-driven testing* begins by picturing the impact a test is intended to have upon all the stakeholders in a society, and work backwards to the actual design of the test (Fulcher and Davidson, 2007). This means that we cannot separate the actual practice of writing tests and assessments – the nuts and bolts of test design and creation – from our values. For teachers and other practitioners, this is liberating. It means that our philosophy and understanding of what is valuable and meaningful in society and education are highly relevant to the tests that we use. We can also see why things happen the way they do. And once we can see this, we can also imagine how they might change for the better.

3. Testing rituals

High-stakes externally mandated tests like the Gaokao are easily distinguishable from classroom assessments by another critical feature: the 'rituality' associated with the activity of testing (further discussed in Chapter 9). As the test marks the culmination of secondary education, it is a 'rite of passage', an event that marks a significant stage in life. It also determines the immediate future, and longer-term prospects, of each test taker. Such events are ritualised, following established practices that endow the activity with special meaning. But the rituals themselves are drawn from the values embedded in the educational and social system, in this case, meritocracy and equality of opportunity. Arriving at a pre-specified place at the same time as others, sitting in a designated seat a regulation distance from other seats, and answering the same questions as other learners in the same time period, are all part of this ritual. This testing practice is designed to enable meritocracy by imposing the same conditions upon all test takers. A *standardised test* is defined by Cohen and Wollack (2006: 358) in the following way:

> *Tests are standardized when the directions, conditions of administration, and scoring are clearly defined and fixed for all examinees, administrations, and forms.*

The principle at stake is that any difference between the score of two individuals should directly reflect their ability upon what is being tested. To put it another way, if two individuals have an equal ability on what is being tested, they should get the same score. If one person gets a higher score because she received more time to take the test, or sat

so close to a more able student that she could copy, the principles of meritocracy and equality of opportunity would be compromised.

In the Gaokao, maintaining the principles is taken extremely seriously. Apart from the normal examination regulations, during the two days of testing building sites are closed, aircraft flight paths are changed to avoid low-flying aircraft disturbing students, and test centres are provided with their own police guard to reduce traffic noise and maintain security over test papers. The cost of these measures is extremely high. However, it is known from research that increased noise during a test can in some circumstances result in reduced scores (Haines *et al.*, 2002; Powers *et al.*, 2002) because it affects concentration. If some test centres are subject to noise levels that other tests centres do not experience, any difference in scores could be a result of noise. In testing jargon the impact of any variable like noise upon test scores is called *construct irrelevant variance*, or the variance in scores that is due to a factor in which we are not at all interested. Another such factor is cheating, and so students are often checked with metal detectors as they enter the examination room to ensure they are not carrying mobile devices or any other information storage equipment. *Invigilation*, or *proctoring*, is carried out with great care, and any case of examination fraud is dealt with harshly.

These rituals are repeated around the world. And the rituals are far from a new invention. China's Imperial Examination System was started in the Sui dynasty of 589–618 AD and only came to an end in 1905. Designed to select the most able to fill posts in the civil service, the examinations were free to enter, and open to anyone who wished to participate. Rules were formulated about leaving one's seat, the impropriety of exchanging or dropping test papers, talking to others during the test, gazing at others, changing seats, disobeying instructions from the invigilator, humming, or submitting incomplete test papers (Miyazaki, 1981: 28). These examinations also instituted the principle that the examiners should not know the identity of the test taker when marking work in order to avoid bias or discrimination (Miyazaki, 1981: 117). All of these ancient practices are features of the ritual of testing that teachers around the world are familiar with today.

4. Unintended consequences

If the consequences of testing are those that we intend, and our intentions are good, all is well. However, it is rarely the case that we can have things all our own way. Whenever tests are used in society, even for well-meaning purposes, there are *unintended consequences*. With high-stakes tests, unintended consequences are likely to be much more severe. Let us consider three unintended consequences of tests like the Gaokao.

Perhaps the most obvious unintended consequence is the fact that many students and teachers cease to study the language, and start to study the test. This is done in the belief that there are test-taking strategies that will raise a score even if ability, knowledge or communication skills have not been improved. The effect of a test on teaching is termed *washback* (discussed at length in Chapter 10). While this can be positive or negative, it is often assumed that teaching to the test is negative. Examples of the nega-

tive washback from high-stakes language tests are provided by Mansell (2007: 83–90) in the context of the United Kingdom's foreign language General Certificate of Secondary Education examinations. These include:

- Memorising unanalysed fragments of text that can be assembled to create a variety of 100-word essays on simple topics.
- Memorising scripted fragments of speech in relation to common oral interview-type questions, and extended chunks for presentation-type tasks.
- Teaching written responses to questions, followed by oral memorisation drills, for all common topics such as 'family and friends', 'holidays' or 'shopping'.

Associated with this kind of teaching is the publication of test preparation materials on an industrial scale, and the growth of private schools that specialise in test preparation. These 'cram schools' claim that they can raise test scores through specialised tuition in short time periods, primarily by practising test-type questions over and over again, and learning test-taking strategies. Parental and peer pressure may make students spend considerable periods of out-of-school time in test preparation classes, the value of which are questionable (see Chapter 10).

Another unintended consequence of high-stakes testing is the possibility of deteriorating health. Longer hours of study without periods of rest and relaxation, or even time to pursue hobbies or extra-curricular activities, can lead to tiredness. Given the pressure to succeed, stress levels can be high, and becoming run-down can add significantly to fears of failure. It is not surprising that this can lead to health problems among a growing percentage of the test-taking population. At its worst, some students become clinically depressed and suicide rates increase.

This is not an isolated problem. Mental health and stress-related illnesses have been reported in many countries with high-stakes standards-based tests for high school students. Suggested solutions have included the introduction of more schools-based assessment, the reduction in length of time spent on formal summative assessment, and a move toward test formats that reduce the overuse of memorisation activities in class. Teachers do not wish to see learners put under the kind of pressure that happens in many modern educational systems; it is therefore incumbent upon teachers to engage with testing systems and those who create them to develop less stressful approaches.

The final example concerns 'test migration'. Universities in China allocate numbers of places in advance to the various provinces of the country, for which the students in those provinces are competing. In rural provinces students have to get higher scores than their urban counterparts to get into top universities. This has led to the phenomenon of 'examinee migration', where families move to provinces where they perceive their children have a better chance of success. Some have used this example of 'unfairness' to call for the abolition of the examination system, but nevertheless it is still seen as 'the least bad method we have' of ensuring fairness (People's Daily Online, 2007). This phenomenon, in a variety of guises, is universal.

'Fairness' is difficult to define, but it is a concept that is conjured up to defend (or

criticise) many uses of tests. Consider, for example, the *standards-based testing* systems that are now operated in many countries around the world. One of the uses of test scores in these systems is to create school *league tables*. The rhetoric associated with the justification of such tables emphasises 'openness' and 'transparency' in the accountability of schools and teachers, and the 'freedom of choice' that parents have to send their children to a successful school. However, in league tables there are some schools that will appear towards the bottom of the table, as well as schools that appear towards the top. It is often the case that those at the bottom are situated in areas where families are from lower socioeconomic groups. The 'catchment area' of the school is such that the children are likely to be those with fewer life opportunities and experiences on purely financial grounds. There is a resulting pressure upon families to move into the catchment areas of the better schools so that their children are more likely to receive what they perceive to be a better education. The additional demand for houses in these areas pushes up the price of housing, thus reinforcing the lack of mobility of poorer families, and the association between income and education (Leech and Campos, 2003).

In these examples I have attempted to show that testing is not just about creating tests to find out what learners know and can do. When testing is practised outside the classroom and leaves the control of the teacher, it is part of the technology of how a society makes decisions about access to scarce resources. The decisions to test, how to test and what to test are all dependent upon our philosophy of society and our view of how individuals should be treated (Fulcher, 2009). Teachers need to become strong advocates for change and for social justice, rather than bystanders to whom testing 'happens'.

5. Testing and society

The defence of high-stakes externally mandated tests is that they provide fairer access to opportunities and resources than any other method that society has yet conceived. The testing system in China was established in order to reduce the power of the aristocracy in civil administration and open it up to talented individuals from whatever background they came. Spolsky (1995: 16–24) has called the testing practices associated with meritocracy the 'Chinese principle'. He shows how the principle affected the whole of European education in the nineteenth century, with a particular focus on language assessment. He shows that tests, or what Edgeworth (1888: 626) called 'a species of sortition', was a better way of sorting people than on the basis of who their parents were. And we are asked to believe that tests remain the best way of making decisions, even if they are imperfect.

But this is not the only position that we can take. Shohamy (2001a) argues that one reason why test takers and teachers dislike tests so much is that they are a means of control. She argues that many governments and ministries of education use tests to implement language policies and force teachers and students to comply. In her analysis, this takes place mostly within systems that have a strongly enforced national curriculum

with summative high-stakes national tests that are used to ensure that the curriculum is delivered as intended. Shohamy is not reticent about passing judgement upon this use of tests:

> *Implementing policy in such ways is based on threats, fear, myths and power, by convincing people that without tests learning will not occur. It is an unethical way of making policy; it is inappropriate to force individuals in a democratic society. Thus, tests are used to manipulate and control education and become the devices through which educational priorities are communicated to principals, teachers and students.* (Shohamy, 2001a: 115)

This view is firmly based in social criticism drawn from Foucault's (1975) book on discipline and punishment, in which he analysed the history of the penal system as a means of state control. The fact that a discussion of testing appears in this context tells us a great deal about Foucault's views. He argued that authority can control individuals and make them do what it wishes through observation and classification. We can illustrate this with reference to Jeremy Bentham's (1787) views on the ideal prison. In this prison there is a guard tower situated in the centre of the prison with the cells arranged in a circle some distance from the tower (see Figure 1.1). No prisoner can see into the

Fig. 1.1. Jeremy Bentham's Panopticon in action. Credit: © Bettmann/Corbis

cell of another prisoner, nor can he see if there is a guard in the tower – but he assumes that he is being watched nevertheless. The guards in the tower, on the other hand, can observe what is happening in every single cell. Foucault takes Bentham's two principles as the basis for his analysis of control in society: that the exercise of power should be visible (always present), but unverifiable (you do not know if you are being watched at any particular moment). The current trend in some countries to cover the streets with closed-circuit television cameras that cannot always be either switched on or monitored is another realisation of the same theory. And in literature the famous novel *Nineteen Eighty-Four* by George Orwell describes a totalitarian state that uses surveillance of this kind to achieve complete control over the activities and beliefs of its citizens. Orwell coined the phrase 'Big Brother is watching you' that has now entered into everyday language.

In what ways might the examination be similar? It is worth listening to Foucault (1975: 184–185) at some length in his own words:

> *The examination combines the techniques of an observing hierarchy and those of a normalizing judgement. It is a normalizing gaze, a surveillance that makes it possible to qualify, to classify and to punish. It establishes over individuals a visibility through which one differentiates them and judges them. That is why, in all the mechanisms of discipline, the examination is highly ritualized. In it are combined the ceremony of power and the form of the experiment, the deployment of force and the establishment of truth. At the heart of the procedures of discipline, it manifests the subjection of those who are perceived as objects and the objectification of those who are subjected. The superimposition of the power relations and knowledge relations assumes in the examination all its visible brilliance … who will write the … history of the 'examination' – its rituals, its methods, its characters and their roles, its play of questions and answers, its systems of marking and classification? For in this slender technique are to be found a whole domain of knowledge, a whole type of power.*

For Foucault, the ritual is not a rite of passage, but a means of subjecting the test takers to the power of those who control the educational system. It is an act of observation, of surveillance, in which the test taker is subjected to the 'normalizing judgement' of those who expect compliance with the knowledge that is valued by the elite. After all, the answers that the test taker provides will be judged, and in order to do well they have to internalise what is considered 'right' by those in power.

How is this achieved? Firstly, of course, what counts as valuable knowledge and as a 'right' answer is externally controlled. The test takers are treated as 'cases' in a large-scale system that collects and analyses data. Each 'case' is documented according to any personal and demographic information that is collected. As the test data involves numbers, it is given the appearance of 'scientific truth' that is rarely questioned, and the objectification of the individual as a case within a system is complete. But do authorities really behave in this way? The evidence suggests that tests have been used as a means of state control over educational systems and individuals for as long as there has been an educational system. And this has not ceased today. Indeed, with the data storage cap-

acity of modern computers, the tendency is for governments to try and keep much more integrated personal data on each individual unless this is curbed by data protection legislation.

If you have been convinced by this argument so far, it would appear that Foucault has turned upside down the argument that tests are the 'least worst' method of being fair.

The natural reaction of most teachers to what Foucault describes, and what some governments try to achieve through the use of tests, ranges from distaste to outrage. In what follows I will attempt to investigate the origin of the distaste and illustrate it through historical example. The reason for this is very simple. When we read about language tests and educational testing more generally today, it tends to wash over us. The context is so well known, the arguments of the education ministers well rehearsed: Foucault would argue that we are desensitised to what is happening to the point that we become an unquestioning part of the system. It is much easier to see issues in examples that are now alien to us because time has lapsed. Once we are aware of these issues, we can problematise them for our own context, and through the process become more vividly aware of what may be happening. Awareness makes it possible for us to consciously avoid the negative uses of tests, and engage practices from design to implementation that encourage positive test use.

6. Historical interlude I

So let us step back into history for a while, and concentrate on the negative uses of tests, before we return to the positive. The first extensive treatment of the role of education in society is found in Plato's *Republic* (1987), written around 360 BC. In this famous text, Plato sets out his vision of the ideal state. It is constructed of three classes: the Guardians or rulers; the auxiliaries or warriors, who protect the state; and the workers, who generate the wealth. For Plato, the survival of the state depends upon its unity, and so the social structure with its three social castes must be maintained. Of course, this means avoiding any change whatsoever. Plato therefore requires that all people 'devote their full energy to the one particular job for which they are naturally suited' so that 'the integrity and unity of both the individual and the state … be preserved' (1987: 190). The role of education is to perpetuate the class structure of society without change. It was therefore seen as essential that individuals should have no personality, no aspirations, no views, other than those invested in them by the state and their position in it. As Popper (2002: 55–56) puts it:

> *The breeding and the education of the auxiliaries and thereby of the ruling class of Plato's best state [are], like their carrying of arms, a class symbol and therefore a class prerogative. And breeding and education are not empty symbols but, like arms, instruments of class rule, and necessary for ensuring the stability of this rule. They are treated by Plato solely from this point of view, i.e. as powerful political weapons as means which are useful for herding the human cattle, and for unifying the ruling class.*

For Plato, testing was an essential part of the educational system that was designed for the preservation of the elite. It allows the rulers to decide what it was necessary to know, or be able to do, to be a ruler. And a centrally controlled curriculum maximises the stability of the system. Only those who are the most successful in the elite will be allowed to rise to the very top. Plato says of potential Guardians: 'we must see how they stand up to hard work and pain and competitive trials … And any Guardian who survives these continuous trials in childhood, youth, and manhood unscathed, shall be given authority in the state … Anyone who fails them we must reject' (Plato, 1987: 180).

This position is profoundly anti-egualitarian and has very little in common with the 'Chinese principle'. And in fact, it also had very little in common with actual education in democratic Athens of the time, as we know from other sources (Fulcher, 2009). However, Plato has had a very significant impact upon education and assessment practices down the ages. For example, one of Hitler's first acts upon coming to power in the 1930s was to take control of the educational system through the centralisation of curriculum, testing, teacher training and certification. The notion that education was about personal growth and development built into the German educational system by von Humboldt (1854) was replaced with the policy 'that people should not have a will of their own and should totally subordinate themselves' (Cecil, 1971: 428). Education and testing became technological tools to enforce compliance with a collectivist philosophy that required absolute acquiescence.

My experience has been that teachers are far from being anti-egualitarian. Being a professional teacher usually carries with it a desire to provide the very best education to all learners, to help each person achieve their full potential. Such a belief is egualitarian, and implies a commitment to individual growth and development. This is also the critical insight of Dewey (1916): that the goal of personal growth implies the freedom to experiment, make inferences and develop critical awareness. As the level of external control increases, it becomes difficult for teachers to see how this goal can be achieved. I believe that it is this fundamental tension between the tendency of external authorities to impose control through tests, and the ethical imperative of teachers to maximise freedom to achieve individual growth, that results in tensions and frustrations. The examples cited above, from Plato and Nazi Germany, are simply extremes. In both cases the role of the teacher is simply to act as an agent of the state. The teacher is disempowered as a stakeholder and an actor in the educational process. The teacher is de-professionalised.

7. The politics of language testing

It is to be hoped that the extreme educational philosophies and practices discussed in the previous section will never be resurrected. However, education and testing still play a significant role in imposing political policies today. This is particularly the case when testing is used as a tool for policy makers to impose systems that emphasise accountability. That is, the policy makers wish to make teachers and schools accountable to them for

their practices. McNamara and Roever (2006: 213) have claimed that 'the politicization of assessment in these ways is perhaps the most striking feature of current developments in language assessment'. Why would policy makers wish to do this? There are two possible reasons, either or both of which may be operating at any given time.

Reason 1: The progress of individual learners is of central importance in education (an assertion with which teachers would agree). In order for each learner to get the very best education they can, information on institutional performance through tests should be publicly available. This freedom of information provides learners with informed choice (an assertion with which teachers may not agree). League tables also show which institutions are failing, and which are succeeding. This allows parents to choose where to send their children. It also enables central authorities to take remedial action; local information on class test performance allows local managers to deal with underperforming teachers (an assertion with which teachers almost always disagree).

The second two assertions in this reason only hold if we believe that the free-market economy extends to education, and that the role of 'managers' is the close monitoring of outcomes (in terms of test scores) against centrally established targets. In managerial systems success and failure must be measurable in ways that can be reported up and down the system. Test scores are the easiest measures of outcomes to aggregate and report, and for schools they represent the 'bottom line' of the balance sheet – investors in this institution need to know what profit they are getting (Mansell, 2007: 7).

Reason 2: Central authorities are concerned with the efficient operation of the economy, and it is essential to produce the human resources required by business. Many states and supranational organisations are concerned that they are in danger of losing ground in the global economy, and one way of measuring potential economic effectiveness is the readiness of the population to contribute to the economy.

This is how governments use the data generated by PISA literacy tests. International comparisons can feed into national economic strategies that include educational policy. This is where language teachers and educational policy makers are most likely to find themselves in disagreement, for it implies a managerial view of language education that measures success for both teachers and learners in financial terms. The following extract from a popular European magazine is an excellent example of the new managerial view of education.

> *Recently, education has been made the subject of public discussion from the point of view of economic usability. It is seen as some important human resource and must contribute to an optimization of location in a global competition as well as the smooth functioning of social partial systems. Whereas education in former times was associated with the development of individuality and reflection, the unfolding of the muse and creativity, the refinement of perception, expression, taste and judgment, the main things today are the acquisition of competence, standardisation and effective educational processes as well as accreditation and evaluation of educational outcomes.* (*Swiss Magazine*; translation provided by the magazine from the original German)