Testing for Language Teachers

Second Edition

Arthur Hughes

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 1989
Second edition 2003

Printed in the United Kingdom at the University Press, Cambridge

Typeface Sabon 10.5/12. *System* QuarkXPress® [OD&I]

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

ISBN 0 521 823250 hardback
ISBN 0 521 484952 paperback

Contents

1 Teaching and testing

Many language teachers harbour a deep mistrust of tests and of testers. The starting point for this book is the admission that this mistrust is frequently well-founded. It cannot be denied that a great deal of language testing is of very poor quality. Too often language tests have a harmful effect on teaching and learning, and fail to measure accurately whatever it is they are intended to measure.

Backwash

The effect of testing on teaching and learning is known as *backwash*, and can be harmful or beneficial. 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 be positively beneficial. I 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 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 beneficial backwash.

Davies (1968:5) once wrote that 'the good test is an obedient servant since it follows and apes the teaching'. I find it difficult to agree, and perhaps today Davies would as well. 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 likely 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 beneficial 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 beneficial backwash on teaching, it would have a much better reputation among teachers. Chapter 6 of this book is devoted to a discussion of how beneficial backwash can be achieved.

One last thing to be said about backwash in the present chapter is that it can be viewed as part of something more general – the *impact* of assessment. 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 (see Further Reading).

Inaccurate tests

The second reason for mistrusting tests is that very often they 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. To return to an earlier example, 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. Professional testers have expended great effort, and not a little money, in attempts to do it, but they have always failed. We may be able to get an approximate measure, but that is all. When testing is carried out on a very large scale, when the scoring of tens of thousands of compositions might not seem to be a practical proposition, it is understandable that potentially greater accuracy is sacrificed for reasons of economy and convenience. But this does not give testing a good name! And it does set a bad example.

While few teachers would wish to follow that particular example in order to test writing ability, the overwhelming practice in large-scale testing of using multiple choice items does lead to imitation in circumstances where such items are not at all appropriate. What is more, the imitation tends to be 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. On 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.

The need for tests

So far this chapter has been concerned with understanding why tests are so mistrusted by many language teachers, and how this mistrust is often justified. One conclusion drawn from this might be that we would be better off without language tests. Teaching is, after all, the primary activity; if testing comes in conflict with it, then it is testing that should go, especially when it has been admitted that so much testing provides inaccurate information. However, information about people's language ability is often very useful and sometimes necessary. It is difficult to imagine, for example, British and American universities accepting students from overseas without some knowledge of their proficiency in English. The same is true for organisations hiring interpreters or translators. They certainly need dependable measures of language ability. Within teaching systems, too, so long as it is thought appropriate for individuals to be given a statement of what they have achieved in a second or foreign language, tests of some kind or another will be needed. They will also be needed in order to provide information about the achievement of groups of learners, without which it is difficult to see how rational educational decisions can be made. While for some purposes teachers' informal assessments of their own students are both appropriate and sufficient, this is not true for the cases just mentioned. Even without considering the possibility of bias, we have to recognise the need for a common yardstick, which tests provide, in order to make meaningful comparisons.

Testing and assessment

The focus of this book is on more or less formal testing. But testing is not, of course, the only way in which information about people's language ability can be gathered. It is just one form of assessment, and other methods will often be more appropriate. It is helpful here to make clear the difference between *formative* and *summative* assessment. Assessment is formative when teachers use it to check on the progress of their students, to see how far they have mastered what they should have learned, and then use this information to modify their future teaching plans. Such assessment can also be the basis for feedback to the students. Informal tests or quizzes may have a part to play in formative assessment but so will simple observation (of performance on learning tasks, for example) and the study of portfolios that students have made of their work. Students themselves may be encouraged to carry out *self-assessment* in order to monitor their progress, and then modify their own learning objectives.

Summative assessment is used at, say, the end of the term, semester, or year in order to measure what has been achieved both by groups and by individuals. Here, for the reasons given in the previous section, formal tests are usually called for. However, the results of such tests should not be looked at in isolation. A complete view of what has been achieved should include information from as many sources as possible. In an ideal world, the different pieces of information from all sources, including formal tests, should be consistent with each other. If they are not, the possible sources of these discrepancies need to be investigated.

What is to be done?

I believe that the teaching profession can make three contributions to the improvement of testing: they can write better tests themselves; they can enlighten other people who are involved in testing processes; and they can put pressure on professional testers and examining boards, to improve *their* tests. This book aims to help them do all three. The first aim is easily understood. One would be surprised if a book with this title did not attempt to help teachers write better tests. The second aim is perhaps less obvious. It is based on the belief 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 I have in mind include test takers, teachers, test writers, school or college administrators, education authorities, examining bodies and testing institutions. 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.

For the reader who doubts the relevance of the third aim, let this chapter end with a further reference to the testing of writing 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 beneficial backwash.

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 beneficial? 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?

Further reading

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* (Volume 14, Number 3) devoted to ethics in language testing. For an early discussion of the ethics of language testing, see Spolsky (1981). The International Language Testing Association has developed a code of ethics (adopted in 2000) which can be downloaded from the Internet (see the book's website). Kunnan (2000) is concerned with fairness and validation in language

testing. Rea-Dickens and Gardner (2000) examine the concept and practice of formative assessment. Alderson and Clapham (1995) make recommendations for classroom assessment. Brown and Hudson (1998) present teachers with alternative ways of assessing language. Nitko (1989) offers advice on the designing of tests which are integrated with instruction. Ross (1998) reviews research into self assessment. DeVicenzi (1995) gives advice to teachers on how to learn from standardised tests. Gipps (1990) and Raven (1991) draw attention to the possible dangers of inappropriate assessment. For an account of how the introduction of a new test can have a striking beneficial effect on teaching and learning, see Hughes (1988a).