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.