1 The shape of things to come: will it be the normal distribution?

Plenary address to the ALTE Conference, Barcelona, July 2001

Charles Alderson
Department of Linguistics and Modern English Language
Lancaster University

Introduction

In this paper I shall survey developments in language testing over the past decade, paying particular attention to new concerns and interests. I shall somewhat rashly venture some predictions about developments in the field over the next decade or so and explore the shape of things to come.

Many people see testing as technical and obsessed with arcane procedures and obscure discussions about analytic methods expressed in alphabet soup, such as IRT, MDS, SEM and DIF. Such discourses and obsessions are alien to teachers, and to many other researchers. In fact these concepts are not irrelevant, because many of them are important factors in an understanding of our constructs – what we are trying to test. The problem is that they are often poorly presented: researchers talking to researchers, without being sensitive to other audiences who are perhaps less obsessed with technical matters. However, I believe that recent developments have seen improved communication between testing specialists and those more generally concerned with language education which has resulted in a better understanding of how testing connects to people's lives.

Much of what follows is not necessarily new, in the sense that the issues have indeed been discussed before, but the difference is that they are now being addressed in a more critical light, with more questioning of assumptions and by undertaking more and better empirical research.

Washback and consequential validity

Washback is a good example of an old concern that has become new. Ten years ago washback was a common concept, and the existence and nature of washback was simply accepted without argument. Tests affect teaching. Bad tests have negative effects on teaching; more modern, good tests will have positive effects; therefore change the test and you will change teaching. I certainly believed that, and have published several articles on the topic. But in the late 1980s and early 1990s, Dianne Wall and I were engaged in a project to investigate washback in Sri Lanka, intended to prove that positive washback had been brought about by a suite of new tests. To our surprise we discovered that things were not so simple. Although we found evidence of the impact of tests on the content of teaching, not all of that impact was positive. Moreover, there was little or no evidence of the impact of the test on how teachers taught – on their methodology. As a result we surveyed the literature to seek for parallels, only to discover that there was virtually no empirical evidence on the matter. We therefore decided to problematise the concept (Alderson and Wall 1993). The rest is history, because washback research quickly took off in a fairly big way.

A number of studies on the topic have been reported in recent years and washback, or more broadly the impact of tests and their consequences on society, has become a major concern. Language testing is increasingly interested in what classrooms look like, what actually happens in class, how teachers prepare for tests and why they do what they do. We now have a fairly good idea of the impact of tests on the content of teaching, but we are less clear about how tests affect teachers' methods. What we do know is that the washback is not uniform. Indeed, it is difficult to predict exactly what teachers will teach, or how teachers will teach. In extreme cases, such as TOEFL test preparation, we know that teachers will tend to use test preparation books, but *how* they use them – and above all *why* they use them in the way they do is still in need of research. In short, washback needs explaining.

There have been fewer studies of what students think, what their test preparation strategies are and why they do what they do, but we are starting to get insights. Watanabe (2001) shows that students prepare in particular for those parts of exams that they perceive to be more difficult, and more discriminating. Conversely, those sections perceived to be easy have less impact on their test preparation practices: far fewer students report preparing for easy or non-discriminating exam sections. However, those students who perceived an exam section to be too difficult did not bother preparing for it at all.

Other studies have turned to innovation theory in order to understand how change occurs and what might be the factors that affect washback (e.g. Wall 1999), and this is a promising area for further research. In short, in order to

understand and explain washback, language testing is engaging with innovation theory, with studies of individual teacher thinking and student motivation, and with investigations of classrooms.

Interestingly, however, washback has not yet been properly researched by testing bodies, who may well not welcome the results. Despite the widely claimed negative washback of TOEFL, the test developer, Educational Testing Service New Jersey, has not to my knowledge funded or engaged in any washback research and the only empirical study I know of into the impact of TOEFL is an unfunded small-scale study in the USA by Hamp-Lyons and myself (Alderson and Hamp-Lyons 1996).

Hopefully, members of ALTE (the Association of Language Testers in Europe) will begin to study the impact of their tests, rather than simply asserting their beneficial impact. After all, many ALTE tests affect high-stakes decisions. I know of no published washback studies among ALTE partners to date, but would be happy to be proved wrong. Certainly I would urge members of ALTE to initiate investigations into the impact of their tests on classrooms, on teachers, on students, and on society more generally.

The results of washback studies will inevitably be painful, not just for test providers but for teachers, too. From the research that has been done to date, it is becoming increasingly clear that a) what teachers say they do is not what they do in class; b) their public reasons for what they do do not always mesh with their real reasons; and c) much of teacher-thinking is vague, muddled, rationalised, prejudiced, or simply uninformed. It is certainly not politically correct to make such statements, and the teacher education literature is full of rosy views of teaching and teachers. But I firmly believe that we need a more realistic, honest view of why teachers do what they do.

Ethics: new focus on old issues

Hamp-Lyons (1997) argues that the notion of washback is too narrow a concept, and should be broadened to cover 'impact' more generally, which she defines as the effect of tests on society at large, not just on individuals or on the educational system. In this, she is expressing a growing concern with the political and related ethical issues that surround test use. Others, like Messick (1994, 1996), have redefined the scope of validity and validation to include what he calls consequential validity – the consequences of test score interpretation and use. Messick also holds that all testing involves making value judgements, and therefore language testing is open to a critical discussion of whose values are being represented and served, which in turn leads to a consideration of ethical conduct.

Tests and examinations have always been used as instruments of social policy and control, with the gate-keeping function of tests often justifying their existence. Davies (1997) argues that language testing is an intrusive

practice, and since tests often have a prescriptive or normative role, then their social consequences are potentially far-reaching. In the light of such impact, he proposes the need for a professional morality among language testers, both to protect the profession's members and to protect the individual within society from misuse and abuse of testing instruments. However, he also argues that the morality argument should not be taken too far, lest it lead to professional paralysis, or cynical manipulation of codes of practice.

A number of case studies illustrate the use and misuse of language tests. Two examples from Australia (Hawthorne 1997) are the use of the access test to regulate the flow of migrants into Australia, and the step test, allegedly designed to play a central role in the determining of asylum seekers' residential status. Similar misuses of the IELTS test to regulate immigration into New Zealand are also discussed in language testing circles – but not yet published in the literature. Perhaps the new concern for ethical conduct will result in more whistle-blowing accounts of such misuse. If not, it is likely to remain so much hot air.

Nevertheless, an important question is: to what extent are testers responsible for the consequences, use and misuse of their instruments? To what extent can test design prevent misuse? The ALTE Code of Practice is interesting, in that it includes a brief discussion of test developers' responsibility to help users to interpret test results correctly, by providing reports of results that describe candidate performance clearly and accurately, and by describing the procedures used to establish pass marks and/or grades. If no pass mark is set, ALTE members are advised to provide information that will help users set pass marks when appropriate, and they should warn users to avoid anticipated misuses of test results.

Despite this laudable advice, the notion of consequential validity is in my view highly problematic because, as washback research has clearly shown, there are many factors that affect the impact a test will have, and how it will be used, misused and abused. Not many of these can be attributed to the test, or to test developers, and we need to demarcate responsibility in these areas. But, of course, the point is well taken that testers should be aware of the consequences of their tests, and should ensure that they at least behave ethically. Part of ethical behaviour, I believe, is indeed investigating, not just asserting, the impact of the tests we develop.

Politics

Clearly, tests can be powerful instruments of educational policy, and are frequently so used. Thus testing can be seen, and increasingly is being seen, as a political activity, and new developments in the field include the relation between testing and politics, and the politics of testing (Shohamy 2001).

But this need not be only at the macro-political level of national or local

government. Politics can also be seen as tactics, intrigue and manoeuvring within institutions that are themselves not political, but rather commercial, financial and educational. Indeed, I argue that politics with a small 'p' includes not only institutional politics, but also personal politics: the motivation of the actors themselves and their agendas (Alderson 1999).

Test development is a complex matter intimately bound up with a myriad of agendas and considerations. Little of this complex interplay of motives and actions surfaces in the language-testing literature (just as so little of teachers' motives for teaching test-preparation lessons the way they do is ever addressed critically in the literature). I do not have the space to explore the depth and breadth of these issues, but I would call for much more systematic study of the true politics of testing.

Clearly, any project involving change on a national level is complex. However, in language testing we often give the impression that all we have to do to improve our tests is to concentrate on the technical aspects of the measuring instruments, design appropriate specifications, commission suitable test tasks, devise suitable procedures for piloting and analysis, train markers, and let the system get on with things. Reform, in short, is considered a technical matter, not a social problem.

However, innovations in examinations are social experiments that are subject to all sorts of forces and vicissitudes, and are driven by personal, institutional, political and cultural agendas, and a concentration on the technical at the expense of these other, more powerful, forces risks the success of the innovation. But to concentrate on the macro-political at the expense of understanding individuals and their agendas is equally misleading. In my experience, the macro-politics are much less important than the private agendas, prejudices and motivations of individuals – an aspect of language testing never discussed in the literature, only in bars on the fringes of meetings and conferences. Exploring this area will be difficult, partly because of the sensitivities involved and partly because there are multiple perspectives on any event, and particularly on political events and actions. It will probably be difficult to publish any account of individual motivations for proposing or resisting test use and misuse. That does not make it any less important.

Testing is crucially affected by politics and testers need to understand matters of innovation and change: how to change, how to ensure that change will be sustainable, how to persuade those likely to be affected by change and how to overcome, or at least understand, resistance.

Standards: codes of practice and levels

Given the importance of tests in society and their role in educational policy, and given recent concerns with ethical behaviour, it is no surprise that one area of increasing concern has been that of standards in testing. One common

meaning of standards is that of 'levels of proficiency'– 'what standard have you reached?' Another meaning is that of procedures for ensuring quality, as in 'codes of practice'.

Language testing has developed a concern to ensure that tests are developed following appropriate professional procedures. Despite the evidence accumulated in the book I co-authored (Alderson, Clapham and Wall 1995), where British EFL exam boards appeared not to feel obliged to follow accepted development procedures or to be accountable to the public for the qualities of the tests they sold, things have now changed, and a good example of this is the publication of the ALTE Code of Practice, which is intended to ensure quality work in test development throughout Europe. 'In order to establish common levels of proficiency, tests must be comparable in terms of quality as well as level, and common standards need, therefore, to be applied to their production.' (ALTE 1998). Mechanisms for monitoring, inspecting or enforcing such a code do not yet exist, and therefore the consumer should still be sceptical, but having a Code of Practice to refer to does strengthen the position of those who believe that testing should be held accountable for its products and procedures.

The other meaning of 'standards', as 'levels of proficiency', has been a concern for some considerable time, but has received new impetus, both with recent changes in Central Europe and with the publication of the Council of Europe's Common European Framework. The Council of Europe's Common European Framework is not only seen as independent of any possible vested interest, it also has a long pedigree, originating over 25 years ago in the development of the Threshold level, and thus its broad acceptability is almost guaranteed. In addition, the development of the scales of various aspects of language proficiency that are associated with the Framework has been extensively researched and validated, by the Swiss Language Portfolio project and DIALANG amongst others. I can confidently predict that we will hear much more about the Common European Framework in the coming years, and that it will increasingly become a point of reference for language examinations across Europe and beyond.

National tests

One of the reasons we will hear a great deal about the Common European Framework in the future is because of the increasing need for mutual recognition and transparency of certificates in Europe, for reasons of educational and employment mobility. National language qualifications, be they provided by the state or by quasi-private organisations, vary enormously in their standards – both quality standards and standards as levels. International comparability of certificates has become an economic as well as an educational imperative, and the availability of a transparent, independent

framework like the Common European Framework is central to the desire to have a common scale of reference and comparison.

In East and Central Europe in particular, there is great interest in the Framework, as educational systems are in the throes of revising their assessment procedures. What is desired for the new reformed exams is that they should have international recognition, unlike the current school-leaving exams which in many places are seen as virtually worthless. Being able to anchor their new tests against the Framework is seen as an essential part of test development work, and there is currently a great deal of activity in the development of school-leaving achievement tests in the region.

National language tests have always been important, of course, and we still see much activity and many publications detailing this work, although unfortunately much of this is either description or heated discussion and is not based on research into the issues.

This contrasts markedly with the literature surrounding international language proficiency examinations, such as TOEFL, TWE, IELTS and some Cambridge exams. Empirical research into various aspects of the validity and reliability of such tests continues apace, often revealing great sophistication in analytic methodology, and such research is, in general, at the leading edge of language-testing research. This, however, masks an old concern: there is a tendency for language-testing researchers to write about large-scale international tests, and not about local achievement tests (including school-leaving tests that are clearly relatively high stakes). Given the amount of language testing that must be going on in the real world, there is a relative dearth of publications and discussions about achievement testing (especially low-stakes testing), and even less about progress testing.

Test development work that is known to be going on, e.g. in Slovakia, the Baltics, St Petersburg and many other places, tends not to get published. Why is this? In many cases, reports are simply not written up, so the testing community does not know about the work. Perhaps those involved have no incentive to write about their work. Or perhaps this is because test developers feel that the international community is not interested in their work, which may not be seen as contributing to debates about test methods, appropriate constructs, the consequences of test use, and so on. However, from my own involvement in exam reform in Hungary, I can say that there is a lot of innovative work that is of interest to the wider community and that should be published. In Hungary we have published articles based on the English examination reform, addressing issues such as the use of sequencing as a test method, research into paired oral tests, and procedures for standard setting, and we have even produced evidence to inform an ongoing debate in Hungary about how many hours per week should be devoted to foreign-language education in the secondary school system.

Indeed, testing is increasingly seen as a means of informing debates in

language education more generally. Examples of this include baseline studies associated with examination reform, which attempt to describe current practice in language classrooms. What such studies have revealed has been used in INSET and PRESET in Central Europe. Washback studies can also be used in teacher training, both in order to influence test preparation practices and also, more generally, to encourage teachers to reflect on the reasons for their and others' practices.

Testing and language education

I am, of course, not the first to advance the argument that testing should be close to – indeed central to – language education. Not only as a means by which data can be generated to illuminate issues in language education, as I have suggested, and not only as an external control of curricular achievement, or as a motivator of students within classrooms. But also, and crucially, as contributing to and furthering language learning. It is a commonplace to say that tests provide essential feedback to teachers on how their learners are progressing, but frankly, few tests do. Teacher-made tests are often poorly designed, provide little meaningful information, and serve more as a disciplinary function than a diagnostic one. Many language textbooks are not accompanied by progress or achievement tests, and those that are are rarely properly piloted and researched.

There is a great lack of interest among testing researchers in improving classroom-based testing. And those who reject testing, as I shall discuss later, claim that teachers know better than tests anyway: they have a more intimate, deep and complex knowledge of what the students they teach are capable of. Frankly I doubt this, and I have yet to see convincing (or indeed *any*) evidence that this might be the case. What language education needs is research and development work aimed at improving regular classroom assessment practice. This can partly be addressed by INSET workshops helping teachers to write better tests, but these can only reach so many teachers, and in any case teachers need more incentives to change their behaviour than can be provided by the occasional workshop.

What holds much more promise is the development of low-stakes tests that can be made available to teachers for little or no charge via the Internet, which do not deliver certificates, but which are deliberately aimed at learning, at supporting teachers' needs for student placement, at the diagnosis of students' strengths and weaknesses, and at assessing student progress. There are already many language tests out there on the Internet, but the quality of many of these is atrocious, and what are urgently needed are high-quality, professionally-developed tests that can be made available to regular classroom teachers to select to suit their own particular needs.

At the centre of testing for learning purposes, however, is the key question:

what CAN we diagnose? Diagnosis is essentially done for individuals, not groups, and testing researchers will increasingly have to ask themselves: what do we understand about individual rather than group performances? Given what we know or suspect about the variation across individuals on tests, what confidence can we have in our knowledge of which ability or process underlies a test taker's response to an item? I shall return to this issue below, but here I raise the question: does it matter if individual learners respond to test items in different ways? If we are dealing with total scores, probably not, because the whole is more than the parts, and normally decisions are made on the basis of total scores, not responses to individual items. But when we are looking at individual skills and individual weaknesses, when we are attempting diagnosis, rather than the characterisation of overall proficiency, what confidence can or must we have that we are accurate? What can we say with confidence about an individual, about his/her individual knowledge or ability, other than through a detailed examination of each item and each response? In the past we could not dream of conducting such a detailed examination on anything other than a very small scale, but now we can. With the help of technology, we can reveal detailed item-level scores and responses (as provided in DIALANG, for example). Thanks to computers we are now able to face the dilemma: what does it all mean?

Technology and testing

Although computers have been used in language testing for a long time, the 1990s saw an explosion of interest in mounting tests on computer, as personal computers and computer labs became much more available, and the accessibility of the World Wide Web increased.

Many have pointed out that computer-based testing relies overwhelmingly on selected response (typically multiple-choice) discrete-point tasks rather than performance-based items, and thus computer-based testing may be restricted to testing linguistic knowledge rather than communicative skills. No doubt this is largely due to the fact that computer-based tests require the computer to score responses.

But recent developments offer some hope. Human-assisted scoring systems (where most scoring of responses is done by computer but responses that the programs are unable to score are given to humans for grading) could reduce this dependency. Free-response scoring tools are capable of scoring responses up to 15 words long, which correlate with human judgements at impressively high levels. ETS has developed 'e-rater' which uses natural language-processing techniques to duplicate the performance of humans rating open-ended essays. Already, the system is used to rate GMAT essays and research is on-going for other programs, including second/foreign language testing situations.

Another example is PhonePass, which is delivered over the telephone, using tasks like reading aloud, repeating sentences, saying opposite words, and giving short answers to questions. Speech synthesis techniques are used to rate the performances, and impressive reliability coefficients have been found as well as correlations with the Test of Spoken English and with interviews.

The advantages of computer-based assessment are already evident, not only in that they can be more user-friendly, but also because they can be more compatible with language pedagogy. Computer-based testing removes the need for fixed delivery dates and locations normally required by traditional paper-and-pencil-based testing. Group administrations are unnecessary, and users can take the test when they wish, and on their own. Whilst diskette- and CD-ROM-based tests also have such advantages, tests delivered over the Internet are even more flexible in this regard: purchase of disks is not required, and anybody with access to the Internet can take a test. Moreover, disks are fixed in format, and once the disk has been created and distributed, it cannot easily be updated. However, with tests delivered by the Internet, access is possible to a much larger database of items, which can be constantly updated. Using the Internet, tests can be piloted alongside live test items. Once a sufficient number of responses has been obtained, they could be calibrated automatically and could then be entered into the live database. Use of the Internet also means that results can be sent immediately to designated score users.

Access to large databases of items means that test security can be greatly enhanced, since tests can be created by randomly accessing items in the database and producing different combinations of items. Thus any one individual is exposed to only a tiny fraction of available items and any compromise of items that might occur will have a negligible effect.

Test results can be made available immediately, unlike paper-and-pencil-based tests, which require time to be collected, marked and for the results to be issued. As well as being of obvious benefit to the users (receiving institutions, as well as candidates), the major pedagogic advantage is that of immediate feedback to the learner, either after each item has been responded to, or at the end of a sub-test, or after the whole test. Feedback given immediately after an activity has been completed is likely to be more meaningful and to have more impact than feedback which is substantially delayed. In traditional paper-and-pencil tests, the test results can be delayed for several months.

If feedback is given immediately after an item has been attempted, users could be allowed to make a second attempt at the item – with or without penalties for doing so in the light of feedback. The interesting question then arises: if the user gets the item right the second time, which is the true measure of ability, the performance before or after the feedback? I would argue that the second performance is a better indication, since it results from the users'