State-of-the-Art Review

Language testing and assessment (Part I)

J Charles Alderson and Jayanti Banerjee Lancaster University, UK

Introduction

This is the third in a series of State-of-the-Art review articles in language testing in this journal, the first having been written by Alan Davies in 1978 and the second by Peter Skehan in 1988/1989. Skehan remarked that testing had witnessed an explosion of interest, research and publications in the ten years since the first review article, and several commentators have since made similar remarks. We can only concur, and for quantitative corroboration would refer the reader to Alderson (1991) and to the International Language Testing Association (ILTA) Bibliography 1990–1999 (Banerjee *et al.*, 1999). In the latter bibliography, there are 866 entries, divided into 15 sections, from Testing Listening to Ethics and Standards. The field has become so large and so active that it is virtually impossible to do justice to it, even in a multi-part 'State-of-the-Art' review like this, and it is changing so rapidly that any prediction of trends is likely to be outdated before it is printed.

In this review, therefore, we not only try to avoid anything other than rather bland predictions, we also acknowledge the partiality of our choice of topics and trends, as well, necessarily, of our selection of publications. We have tried to represent the field fairly, but have tended to concentrate on articles rather than books, on the grounds that these are more likely to reflect the state of the art than are full-length books. We have also referred to other similar reviews published in the last 10 years or so, where we judged it relevant. We have usually begun our review with articles printed in or around 1988, the date of the last review, aware that this is now 13 years ago, but also conscious of the need to cover the period since the last major review in this journal. However, we have also, where we felt it appropriate, included articles published somewhat earlier.

This review is divided into two parts, each of roughly equal length. The bibliography for works referred to in each part is published with the relevant part, rather than in a complete bibliography at the end. Therefore, readers wishing to have a complete bibliography will have to put both parts together.

The rationale for the organisation of this review is that we wished to start with a relatively new concern in language testing, at least as far as publication of empirical research is concerned, before moving on to more traditional ongoing concerns and ending with aspects of testing not often addressed in international reviews, and remaining problems. Thus, we begin with an account of research into washback, which then leads us to ethics, politics and standards. We then examine trends in testing on a national level, followed by testing for specific purposes. Next, we survey developments in computer-based testing before moving on to look at self-assessment and alternative assessment. Finally in this first part, we survey a relatively new area: the assessment of young learners.

In the second part, we address new concerns in test validity theory, which argues for the inclusion of test consequences in what is now generally referred to as a unified theory of construct validity. Thereafter we deal with issues in test validation and test development, and examine in some detail more traditional research into the nature of the constructs (reading, listening, grammatical abilities, etc.) that underlie tests. Finally we discuss a number of remaining controversies and puzzles that we call, following McNamara (1995), 'Pandora's Boxes'.

We are very grateful to many colleagues for their assistance in helping us draw up this review, but in particular we would like to acknowledge the help, advice and support of the Lancaster Language Testing Research Group, above all of Dianne Wall and Caroline Clapham, for their invaluable and insightful comments. All faults that remain are entirely our responsibility.

J Charles Alderson is Professor of Linguisics and English Language Education at Lancaster University. He holds an MA in German and French from Oxford University and a PhD in Applied Linguistics from Edinburgh University. He is co-editor of the journal Language Testing *(Edward Arnold), and co-editor of the* Cambridge Language Assessment Series *(C.U.P.), and has published many books and articles on language testing, reading in a foreign language, and evaluation of language education.*

Jayanti Banerjee is a PhD student in the Department of Linguistics and Modern English Language at Lancaster University. She has been involved in a number of test development and research projects and has taught on introductory testing courses. She has also been involved in teaching English for Academic Purposes (EAP) at Lancaster University. Her research interests include the teaching and assessment of EAP as well as qualitative research methods. She is particularly interested in issues related to the interpretation and use of test scores.

Washback

The term 'washback' refers to the impact that tests have on teaching and learning. Such impact is usually seen as being negative: tests are said to force teachers to do things they do not necessarily wish to do. However, some have argued that tests are potentially also 'levers for change' in language education: the argument being that if a bad test has negative impact, a good test should or could have positive washback (Alderson, 1986b; Pearson, 1988).

Interestingly, Skehan, in the last review of the State of the Art in Language Testing (Skehan, 1988, 1989), makes only fleeting reference to washback, and even then, only to assertions that communicative language testing and criterion-referenced testing are likely to lead to better washback – with no evidence cited. Nor is research into washback signalled as a likely important future development within the language testing field. Let those who predict future trends do so at their peril!

In the *Annual Review of Applied Linguistics* series, equally, the only substantial reference to washback is by McNamara (1998) in a chapter entitled: 'Policy and social considerations in language assessment'. Even the chapter entitled 'Developments in language testing' by Douglas (1995) makes no reference to washback. Given the importance assigned to consequential validity and issues of consequences in the general assessment literature, especially since the popularisation of the Messickian view of an all-encompassing construct validity (see Part Two), this is remarkable, and shows how much the field has changed in the last six or seven years. However, a recent review of validity theory (Chapelle, 1999) makes some reference to washback under construct validity, reflecting the increased interest in the topic.

Although the notion that tests have impact on teaching and learning has a long history, there was surprisingly little empirical evidence to support such notions until recently. Alderson and Wall (1993) were among the first to problematise the notion of test washback in language education, and to call for research into the impact of tests. They list a number of 'Washback Hypotheses' in an attempt to develop a research agenda. One Washback Hypothesis, for example, is that tests will have washback on *what* teachers teach (the content agenda), whereas a separate washback hypothesis might posit that tests also have impact on *how* teachers teach (the methodology agenda). Alderson and Wall also hypothesise that high-stakes tests – tests with important consequences – would have more impact than low-stakes tests. They urge researchers to broaden the scope of their enquiry, to include not only attitude measurement and teachers' accounts of washback but also classsroom observation. They argue that the study of washback would benefit from a better understanding of student motivation and of the nature of innovation in education, since the notion that tests will automatically have an impact on the curriculum and on learning has been advocated atheoretically. Following on from this suggestion, Wall (1996) reviews key concepts in the field of educational innovation and shows how they might be relevant to an understanding of whether and how tests have washback. Lynch and Davidson (1994) describe an approach to criterion-referenced testing which involves practising teachers in the translation of curricular goals into test specifications. They claim that this approach can provide a link between the curriculum, teacher experience and tests and can therefore, presumably, improve the impact of tests on teaching.

Recently, a number of empirical washback studies have been carried out (see, for example, Khaniyah, 1990a, 1990b; Shohamy, 1993; Shohamy *et al.*, 1996; Wall & Alderson, 1993; Watanabe, 1996; Cheng, 1997) in a variety of settings. There is general agreement among these that high-stakes tests do indeed impact on the content of teaching and on the nature of the teaching materials. However, the evidence that they impact on how teachers teach is much scarcer and more complicated. Wall and Alderson (1993) found no evidence for any change in teachers' methodologies before and after the introduction of a new style school-leaving examination in English in Sri Lanka. Alderson and Hamp-Lyons (1996) show that teachers may indeed change the way they teach when teaching towards a test (in this case, the TOEFL – Test of English as a Foreign Language), but they also show that the nature of the change and the methodology adopted varies from teacher to teacher, a conclusion supported by Watanabe's 1996 findings. Alderson and Hamp-Lyons argue that it is not enough to describe whether and how teachers might adapt their teaching and the content of their teaching to suit the test. They believe that it is important to explain *why* teachers do what they do, if we are to understand the washback effect. Alderson (1998) suggests that testing researchers should explore the literature on teacher cognition and teacher thinking to understand better what motivates teacher behaviour. Cheng (1997) shows that teachers only adapt their methodology slowly, reluctantly and with difficulty, and suggests that this may relate to the constraints on teachers and teaching from the educational system generally. Shohamy *et al.* (1996) show that the nature of washback varies according to factors such as the status of the language being tested, and the uses of the test. In short, the phenomenon of washback is slowly coming to be recognised as a complex matter, influenced by many factors other than simply the existence of a test or the nature of that test. Nevertheless, no major studies have yet been carried out into the effect of test preparation on test performance, which is remarkable, given the prevalence, for high-stakes tests at least, of test preparation courses.

Hahn *et al.* (1989) conducted a small-scale study

of the effects on beginning students of German of whether they were or were not graded on their oral performance in the first six months of instruction. Although no effects on developing oral proficiency were found, attitudes in the two groups were different: those who had been graded considered the experience stressful and unproductive, whereas the group that had not been graded would like to have been graded. Moeller and Reschke (1993) also found no effect whatsoever of the formal scoring of classroom performance on student proficiency or achievement. More studies are needed of learners' views of tests and test preparation.

There are in fact remarkably few studies of the impact of tests on motivation or of motivation on test preparation or test performance. A recent exception is Watanabe (2001). Watanabe calls his study a hypothesis-generating exercise, acknowledging that the relationship between motivation and test preparation is likely to be complex. He interviewed Japanese university students about their test preparation practices. He found that attitudes to test preparation varied and that impact was far from uniform, although those exams which the students thought most important for their future university careers usually had more impact than those perceived as less critical. Thus, if an examination for a university which was the student's first choice contained grammar-translation tasks, the students reported that they had studied grammar-translation exercises, whereas if a similar examination was offered by a university which was their second choice, they were much less likely to study translation exercises. Interestingly, students studied in particular those parts of the exam that they perceived to be more difficult, and more discriminating. Conversely those sections perceived to be easy had less impact on their test preparation practices: far fewer students reported 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. Watanabe concludes that washback is caused by the interplay between the test and the test taker in a complex manner, and he emphasises that what may be most important is not the objective difficulty of the test, but the students' perception of difficulty.

Wall (2000) provides a very useful overview and up-date of studies of the impact of tests on teaching, from the field of general education as well as in language education. She summarises research findings which show that test design is only one of the factors affecting washback, and lists as factors influencing the nature of test washback:

> teacher ability, teacher understanding of the test and the approach it was based on, classroom conditions, lack of resources, management practices within the school... the status of the subject within the curriculum, feedback mechanisms between the schools and the testing agency, teacher style, commitment and willingness to innovate, teacher background, the general social and political context, the time that has elapsed since the test was introduced, and the role of publishers in materials design and teacher training (2000: 502).

In other words, test washback is far from being simply a technical matter of design and format, and needs to be understood within a much broader framework. Wall suggests that such a framework might usefully come from studies and theories of educational change and innovation, and she summarises the most important findings from these areas.

She develops a framework derived from Henrichsen (1989), and owing something to the work of Hughes (1993) and Bailey (1996), and shows how such a framework might be applied to understanding better the causes and nature of washback. She makes a number of recommendations about the steps that test developers might take in the future in order to assess the amount of risk involved in attempting to bring about change through testing. These include assessing the feasibility of examination reform by studying the 'antecedent' conditions – what is increasingly referred to as a 'baseline study' (Weir & Roberts, 1994, Fekete *et al.*, 1999); involving teachers at all stages of test development; ensuring the participation of other key stakeholders including policy-makers and key institutions; ensuring clarity and acceptability of test specifications, and clear exemplification of tests, tasks, and scoring criteria; full piloting of tests before implementation; regular monitoring and evaluation not only of test performance but also of classrooms; and an understanding that change takes time. Innovating through tests is not a quick fix if it is to be beneficial. 'Policy makers and test designers should not expect significant impact to occur immediately or in the form they intend. They should be aware that tests on their own will not have positive impact if the materials and practices they are based on have not been effective. They may, however, have negative impact and the situation must be monitored continuously to allow early intervention if it takes an undesirable turn' (2000: 507).

Similar considerations of the potential complexity of the impact of tests on teaching and learning should also inform research into the washback of existing tests. Clearly this is a rich field for further investigation. More sophisticated conceptual frameworks, which are slowly developing in the light of research findings and related studies into innovation, motivation theory and teacher thinking, are likely to provide better understanding of the reasons for washback and an explanation of how tests might be developed to contribute to the engineering of desirable change.

Ethics in language testing

Whilst Alderson (1997) and others have argued that testers have long been concerned with matters of

fairness (as expressed in their ongoing interest in validity and reliability), and that striving for fairness is an aspect of ethical behaviour, others have separated the issue of ethics from validity, as an essential part of the professionalising of language testing as a discipline (Davies, 1997). Messick (1994) argues 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; this in turn leads to a consideration of ethical conduct. Messick (1994, 1996) has redefined the scope of validity to include what he calls consequential validity – the consequences of test score interpretation and use. Hamp-Lyons (1997) argues that the notion of washback is too narrow and should be broadened to cover 'impact', defined as the effect of tests on society at large, not just on individuals or on the educational system. In this, she is expressing a concern that has grown in recent years with the political and related ethical issues which surround test use.

Both McNamara (1998) and Hamp-Lyons (1998) survey the emerging literature on the topic of ethics, and highlight the need for the development of language testing standards (see below). Both comment on a draft Code of Practice sponsored by the International Language Testing Association (ILTA, 1997), but where Hamp-Lyons sees it as a possible way forward, McNamara is more critical of what he calls its conservatism, and this inadequate acknowledgement of the force of current debates on the ethics of language testing. Davies (1997) argues that, since tests often have a prescriptive or normative role, their social consequences are potentially far-reaching. He argues for a professional morality among language testers, both to protect the profession's members, and to protect individuals from the misuse and abuse of tests. 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.

Spolsky (1997) points out that 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. Shohamy (1997a) claims that language tests which contain content or employ methods which are not fair to all test-takers are not ethical, and discusses ways of reducing various sources of unfairness. She also argues that uses of tests which exercise control and manipulate stakeholders rather than providing information on proficiency levels are also unethical, and she advocates the development of 'critical language testing' (Shohamy, 1997b). She urges testers to exercise vigilance to ensure that the tests they develop are fair and democratic, however that may be defined. Lynch (1997) also argues for an ethical approach to language testing and Rea-Dickins (1997) claims that taking full account of the views and interests of various stakeholder groups can democratise the testing process, promote fairness and therefore enhance an ethical approach.

A number of case studies have been presented recently which illustrate the use and misuse of language tests. Hawthorne (1997) describes two examples of the misuse of language tests: 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. Unpublished language testing lore has many other examples, such as the misuse of the General Training component of the International English Language Testing System (IELTS) test with applicants for immigration to New Zealand, and the use of the TOEFL test and other proficiency tests to measure achievement and growth in instructional programmes (Alderson, 2001a). It is to be hoped that the new concern for ethical conduct will result in more accounts of such misuse.

Norton and Starfield (1997) claim, on the basis of a case study in South Africa, that unethical conduct is evident when second language students' academic writing is implicitly evaluated on linguistic grounds whilst ostensibly being assessed for the examinees' understanding of an academic subject. They argue that criteria for assessment should be made explicit and public if testers are to behave ethically. Elder (1997) investigates test bias, arguing that statistical procedures used to detect bias such as DIF (Differential Item Functioning) are not neutral since they do not question whether the criterion used to make group comparisons is fair and value-free. However, in her own study she concludes that what may appear to be bias may actually be construct-relevant variance, in that it indicates real differences in the ability being measured. One similar study was Chen and Henning (1985), who compared international students' performance on the UCLA (University of California, Los Angeles) English as a Second Language Placement Test, and discovered that a number of items were biased in favour of Spanish-speaking students and against Chinese-speaking students. The authors argue, however, that this 'bias' is relevant to the construct since Spanish is much closer to English typologically and therefore biased in favour of speakers of Spanish, who would be expected to find many aspects of English much easier to learn than speakers of Chinese would.

Reflecting this concern for ethical test use, Cumming (1995) reviews the use in four Canadian settings of assessment instruments to monitor learners' achievements or the effectiveness of programmes, and concludes that this is a misuse of such instruments, which should be used mainly for placing students onto programmes. Cumming (1994) asks whether use of language assessment instruments for immigrants to Canada facilitates their successful participation in Canadian society. He argues that

such a criterion should be used to evaluate whether assessment practices are able to overcome institutional or systemic barriers that immigrants may encounter, to account for the quality of language use that may be fundamental to specific aspects of Canadian life, and to prompt majority populations and instruments to better accommodate minority populations.

In the academic context, Pugsley (1988) problematises the assessment of the need of international students for pre- and in-sessional linguistic training in the light of test results. Decisions on whether a student should receive the benefit of additional language instruction are frequently made at the last minute, and in the light of competing demands on the student and on finance. Language training may be the victim of reduced funding, and many academics downplay on importance of language in academic performance. Often, teachers and students perceive students' language related problems differently, and the question of the relevance or influence of the test result is then raised.

In another investigation of score interpretation and use, Yule (1990) analyses the performance of international teaching assistants, attempting to predict on the basis of TOEFL and Graduate Record Examinations Program scores whether the subjects should have received positive or negative recommendations to be teaching assistants. Students who received negative recommendations did indeed have lower scores on both tests than those with positive recommendations, but the relationship between subsequent grade point average (GPA) and positive recommendations only held during the first year of graduate study, not thereafter. The implications for making decisions about the award of teaching assistantships are discussed, and there are obvious ethical implications about the length of time a test score should be considered to be valid.

Both these case studies show the difficulty in interpreting language test results, and the complexity of the issues that surround gate-keeping decisions. They also emphasise that there must be a limit on what information one can ethically expect a language test to deliver, and what decisions test results can possibly inform.

Partly as a result of this heightened interest in ethics and the role of tests in society, McNamara (1998: 313) anticipates in the future:

1. a renewed awareness ... of the socially constructed nature of test performance and test score interpretation;
2. an awareness of the issues raised for testing in the context of English as an International Language;
3. a reconsideration of the social impact of technology in the delivery of tests;
4. an explicit consideration of issues of fairness at every stage of the language testing cycle, and
5. an expanded agenda for research on fairness accompanying test development.

He concludes that we are likely to see 'a broadening of the range of issues involved in language testing research, drawing on, at least, the following disciplines and fields: philosophy, especially ethics and the epistemology of social science; critical theory; policy analysis; program evaluation, and innovation theory' (loc cit).

The International Language Testing Association (ILTA) has recently developed a Code of Ethics (rather than finalising the draft Code of Practice referred to above), which is 'a set of principles which draws upon moral philosophy and strives to guide good professional conduct ... All professional codes should inform professional conscience and judgement ... Language testers are independent moral agents, and they are morally entitled to refuse to participate in procedures which would violate personal moral belief. Language testers accepting employment positions where they foresee they may be called on to be involved in situations at variance with their beliefs have a responsibility to acquaint their employer or prospective employer with this fact. Employers and colleagues have a responsibility to ensure that such language testers are not discriminated against in their workplace.' [http://www.surrey.ac.uk/ELI/ltrfile/ltrframe.html]

These are indeed fine words and the moral tone and intent of this Code is clear: testers should follow ethical practices, and have a moral responsibility to do so. Whether this Code of Ethics will be acceptable in the diverse environments in which language testers work around the world remains to be seen. Some might even see this as the imposition of Western cultural or even political values.

Politics

Tests are frequently used as instruments of educational policy, and they can be very powerful – as attested by Shohamy (2001a). Inevitably, therefore, testing – especially high-stakes testing – is a political activity, and recent publications in language testing have begun to address the relation between testing and politics, and the politics of testing, perhaps rather belatedly, given the tradition in educational assessment in general.

Brindley (1998, 2001) describes the political use of test-based assessment for reasons of public accountabilty, often in the context of national frameworks, standards or benchmarking. However, he points out that political rather than professional concerns are usually behind such initiatives, and are often in conflict with the desire for formative assessment to be closely related to the learning process. He addresses a number of political as well as technical and practical issues in the use of outcomes-based assessment for accountability purposes, and argues for the need for increased consultation between politicians and professionals and for research into the quality of associated instruments.

Politics can be defined as action, or activities, to

achieve power or to use power, and as beliefs about government, attitudes to power, and to the use of power. But this need not only be at the macro-political level of national or local government. National educational policy often involves innovations in testing in order to influence the curriculum, or in order to open up or restrict access to education and employment – and even, as we have seen in the cases of Australia and New Zealand, to influence immigration opportunities. But politics can also operate at lower levels, and can be a very important influence on test development and deployment. Politics can be seen as methods, tactics, intrigue, manoeuvring, within institutions which are themselves not political, but commercial, financial and educational. Indeed, Alderson (1999) argues that politics with a small 'p' includes not only institutional politics, but also personal politics: the motivation of the actors themselves and their agendas. And personal politics can influence both test development and test use.

Experience shows that, in most institutions, test development is a complex matter where individual and institutional motives interact and are interwoven. Yet the language testing literature has virtually never addressed such matters, until very recently. The literature, when it deals with test development matters at all, which is not very often, gives the impression that testing is basically a technical matter, concerned with the development of appropriate specifications, the creation and revision of appropriate test tasks and scoring criteria, and the analysis of results from piloting. But behind that facade is a complex interplay of personalities, of institutional agendas, and of intrigue. Although the macro-political level of testing is certainly important, one also needs to understand individual agendas, prejudices and motivations. However, this is an aspect of language testing which rarely sees the light of day, and which is part of the folklore of language testing.

Exploring such issues is difficult because of the sensitivities involved, and it is difficult to publish any account of individual motivations for proposing or resisting test use and misuse. However, that does not make it any the less important. Alderson (2001a) has the title: 'Testing is too important to be left to testers', and he argues that language testers need to take account of the different perspectives of various stakeholders: not only classroom teachers, who are all too often left out of consideration in test development, but also educational policy makers and politicians more generally. Although there are virtually no studies in this area at present (exceptions being Alderson *et al.*, 2000a, Alderson, 1999, 2001b, and Shohamy, 2001), it is to be hoped that the next decade will see such matters discussed much more openly in language testing, since politics, ethics and fairness are rather closely related. Shohamy (2001b) describes and discusses the potential abuse of tests as instruments of power by authoritarian agencies, and argues for more democratic and accountable testing practice.

As an example of the influence of politics, it is instructive to consider Alderson (2001b). In Hungary translation is still used as a testing technique in the current school-leaving exams, and in the tests administered by the State Foreign Language Examinations Board (SFLEB), a quasi-commercial concern. Language teachers have long expressed their concern at the continued use of a test method which has uncertain validity (this has not been established to date in Hungary), where the marking of translations is felt to be subjective and highly variable, where no marking criteria or scales exist, and where the washback effect is felt to be negative (Fekete *et al.*, 1999). New school-leaving examinations are due to be introduced in 2005, and the intention is not to use translation as a test method in future. However, many people, including teachers, and also Ministry officials, have resisted such a proposal, and it has recently been declared that the Minister himself will take the decision on this matter. Yet the Minister is not a language expert, knows nothing about language testing, and is therefore not technically competent to judge. Many suspect that the SFLEB, which wishes to retain translation, is lobbying the Minister to insist that translation be retained as a test method. Furthermore, many suspect that the SFLEB fears that foreign language examinations, which necessarily do not use translation as a test method, might take over the language test market in Hungary if translation is no longer required (by law) as a testing technique. Alderson (2001b) suggests that translation may be being used as a weapon in the cause of commercial protectionism.

Standards in testing

One area of increasing concern in language testing has been that of standards. The word 'standards' has various meanings in the literature, as the Task Force on Language Testing Standards set up by ILTA discovered (http://www.surrey.ac.uk/ELI/ilta/tfts_report.pdf). One common meaning used by respondents to the ILTA survey was that of procedures for ensuring quality, standards to be upheld or adhered to, as in 'codes of practice'. A second meaning was that of 'levels of proficiency' – 'what standard have you reached?' A related, third meaning was that contained in the phrase 'standardised test', which typically means a test whose difficulty level is known, which has been adequately piloted and analysed, the results of which can be compared with those of a norming population: standardised tests are typically norm-referenced tests. In the latter context 'standards' is equivalent to 'norms'.

In recent years, language testing has sought to establish standards in the first sense (codes of practice) and to investigate whether tests are developed

following appropriate professional procedures. Groot (1990) argues that the standardisation of procedures for test construction and validation is crucial to the comparability and exchangeability of test results across different education settings. Alderson and Buck (1993) and Alderson *et al.* (1995) describe widely accepted procedures for test development and report on a survey of the practice of British EFL examining boards. The results showed that current (in the early 1990s) practice was wanting. Practice and procedures among boards varied greatly, yet (unpublished) information was available which could have attested to the quality of examinations. Exam boards appeared not to feel obliged to follow or indeed to understand accepted procedures, nor did they appear to be accountable to the public for the quality of the tests they produced. Fulcher and Bamford (1996) argue that testing bodies in the USA conduct and report reliability and validity studies partly because of a legal requirement to ensure that all tests meet technical standards. They conclude that British examination boards should be subject to similar pressures of litigation on the grounds that their tests are unreliable, invalid or biased. In the German context, Kieweg (1999) makes a plea for common standards in examining EFL, claiming that within schools there is little or no discussion of appropriate methods of testing or of procedures for ensuring the quality of language tests.

Possibly as a result of such pressures and publications, things appear to be changing in Europe, an example of this being the publication of the ALTE (Association of Language Testers in Europe) 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). To date, no mechanism exists for monitoring whether such standards are indeed being applied, but the mere existence of such a Code of Practice is a step forward in establishing the public accountability of test developers. Examples of how such standards are applied in practice are unfortunately rare, one exception being Alderson *et al.* (2000a), which presents an account of the development of new school-leaving examinations in Hungary.

Work on standards in the third sense, namely 'norms' for different test populations, was less commonly published in the last decade. Baker (1988) discusses the problems and procedures of producing test norms for bilingual school populations, challenging the usual *a priori* procedure of classifying populations into mother tongue and second language groups. Employing a range of statistical measures, Davidson (1994) examines the appropriacy of the use of a nationally standardised test normed on native English speakers, when used with non-English speaking students. Although he concludes that such a use of the test might be defensible statistically, additional measures might nevertheless be necessary for a population different from the norming group.

The meaning of 'standards' as 'levels of proficiency' or 'levels certified by public examinations' has been an issue for some considerable time, but has received new impetus, both with recent developments in Central Europe and with the publication of the Council of Europe's Common European Framework (Council of Europe, 2001). Work in the 1980s by West and Carroll led to the development of the English Speaking Union's Framework (Carroll & West, 1989), but this was not widely accepted, probably because of commercial rivalries within the British EFL examining industry. Milanovic (1995) reports on work towards the establishment of common levels of proficiency by ALTE, which has developed its own definitions of five levels of proficiency, based upon an inspection and comparison of the examinations of its members. This has had more acceptability, possibly because it was developed *by* **cooperating** examination bodies, rather than *for* **competing** bodies. However, such a framework of levels is still not seen by many as being neutral: it is, after all, associated with the main European commercial language test providers. The Council of Europe's Common European Framework, on the other hand, 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 (van Ek, 1977), and thus broad acceptability across Europe is guaranteed. In addition, the scales of various aspects of language proficiency that are associated with the Framework have been extensively researched and validated by the Swiss Language Portfolio Project (North & Schneider, 1998).

de Jong (1992) predicted that international standards for language tests and assessment procedures, and internationally interpretable standards of proficiency would be developed, with the effect that internationally comparable language tests would be established. In the 21st century, that prediction is coming true. It is now clear that the Common European Framework will become increasingly influential because of the growing need for international recognition of certificates in Europe, in order to guarantee educational and employment mobility. National language qualifications, be they provided by the state or by quasi-private organisations, presently vary in their standards – both quality standards and standards as levels. Yet international comparability of certificates has become an economic as well as an educational imperative, especially after the Bologna Declaration of 1999 (http://europa.eu.int/comm/education/socrates/erasmus/bologna.pdf), and the availability of a transparent, independent framework like the Common European framework is crucial to

the attempt to establish a common scale of reference and comparison. Moreover, the Framework is not just a set of scales, it is also a compendium of what is known about language learning, language use and language proficiency. As an essential guide to syllabus construction, as well as to the development of test specifications and rating criteria, it is bound to be used for materials design and textbook production, as well as in teacher education. The Framework is also the anchor point for the European Language Portfolio, and for new diagnostic tests like DIALANG (see below).

The Framework is particularly relevant to countries in East and Central Europe, where many educational systems are currently revising their assessment procedures. The intention is that the reformed examinations should have international recognition, unlike the current school-leaving exams. Calibrating the new tests against the Framework is essential, and there is currently a great deal of activity in the development of school-leaving achievement tests in the region (for one account of such development, see Alderson *et al.*, 2000a). We are confident that we will hear much more about the Common European Framework in the coming years, and it will increasingly become a point of reference for language examinations across Europe and beyond.

National tests

The development of national language tests continues to be the focus of many publications, although many are either simply descriptions of test development or discussions of controversies, rather than reports on research done in connection with test development.

In the UK context, Neil (1989) discusses what should be included in an assessment system for foreign languages in the UK secondary system but reports no research. Roy (1988) claims that writing tasks for modern languages should be more relevant, task-based and authentic, yet criticises an emphasis on letter writing, and argues for other forms of writing, like paragraph writing. Again, no research is reported. Page (1993) discusses the value and validity of having test questions and rubrics in the target language and asserts that the authenticity of such tasks is in doubt. He argues that the use of the target language in questions makes it more difficult to sample the syllabus adequately, and claims that the more communicative and authentic the tasks in examinations become, the more English (the mother tongue) has to be used on the examination paper in order to safeguard both the validity and the authenticity of the task. No empirical research into this issue is reported. Richards and Chambers (1996) and Chambers and Richards (1992) examine the reliability and validity of teacher assessments in oral production tasks in the school-leaving GCSE (General Certificate in Secondary Education) French examination, and find problems particularly in the rating critera, which they hold should be based on a principled model of language proficiency and be informed by an analysis of communicative development. Hurman (1990) is similarly critical of the imprecise specifications of objectives, tasks and criteria for assessing speaking ability in French at GCSE level. Barnes and Pomfrett (1998) find that teachers need training in order to conform to good practice in assessing German for pupils at Key Stage 3 (age 14). Buckby (1999) reports an empirical comparison of recent and older GCSE examinations, to determine whether standards of achievement are falling, and concludes that although the evidence is that standards are indeed being maintained, there is a need for a range of different question types in order to enable candidates to demonstrate their competencies. Barnes *et al.* (1999) consider the recent introduction of the use of bilingual dictionaries in school examinations, report teachers' positive reactions to this innovation, but call for more research into the use and impact of dictionaries on pupil performance in examinations.

Similar research in the Netherlands (Jansen & Peer, 1999) reports a study of the recently introduced use of dictionaries in Dutch foreign language examinations and shows that dictionary use does not have any significant effect on test scores. Nevertheless, pupils are very positive about being allowed to use dictionaries, claiming that it reduces anxiety and enhances their text comprehension. Also in the Netherlands, Welling-Slootmaekers (1999) describes the introduction of a range of open-ended questions into national examinations of reading ability in foreign languages, arguing that these will improve the assessment of language ability (the questions are to be answered in Dutch, not the target foreign language). van Elmpt and Loonen (1998) question the assumption that answering test questions in the target language is a handicap, and report research that shows results to be similar, regardless of whether candidates answered comprehension questions in Dutch (the mother tongue) or in English (the target language). However, Bhgel and Leijn (1999) report research that showed low interrater reliability in marking these new item types and they call for improved assessment practice.

Guillon (1997) evaluates the assessment of English in French secondary schools, criticises the time taken by test-based assessment and the technical quality of the tests, and makes suggestions for improved pupil profiling. Mundzeck (1993) similarly criticises many of the objective tests in use in Germany for official school assessment of modern languages, arguing that they do not reflect the communicative approach to language required by the syllabus. He recommends that more open-ended tasks be used, and that teachers be trained in the reliable use of

valid criteria for subjective marking, instead of their current practice of merely counting errors in production. Kieweg (1992) makes proposals for the improvement of English assessment in German schools, and for the comparability of standards within and across schools.

Dollerup *et al.* (1994) describe the development in Denmark of an English language reading proficiency test which is claimed to help diagnose reading weaknesses in undergraduates. Further afield, in Australia, Liddicoat (1996) describes the Language Profile oral interaction component which sees listening and speaking as interdependent skills and assesses school pupils' ability to participate successfully in spontaneous conversation. Liddicoat (1998) criticises the Australian Capital Territory's guidelines for the assessment of proficiency in languages like Chinese, Japanese and Indonesian, as well as French, German, Spanish and Italian. He argues that empirically-based descriptions of the achievement of learners of such different languages should inform the revision of the descriptors of different levels in profiles of achievement.

In Hong Kong, dissatisfaction with graduating students' levels of language proficiency has resulted in plans for tertiary institution exit controls of language. Li (1997) describes these plans and discusses a range of problematic issues that need resolving before valid measures can be introduced. Coniam (1994, 1995) describes the construction of a common scale which attempts to cover the range of English language ability of Hong Kong secondary school pupils in English. An Item Response Theory-based test bank – the TeleNex – has been constructed to provide teachers both with reference points for ability levels and help in school-based testing.

A similar concern with levels or standards of proficiency is evinced by Peirce and Stewart (1997), who describe the development of the Canadian Language Benchmarks Assessment (CLBA), which is intended to be used across Canada to place newcomers into appropriate English instructional programmes, as part of a movement to establish a common framework for the description of adult ESL language proficiency. The authors give an account of the history of the project and the development of the instruments. However, Rossiter and Pawlikowsska-Smith (1999) are critical of the usefulness of the CLBA because it is based on very broad-band differences in proficiency among individuals and is insensitive to smaller, but important, differences in proficiency. They argue that the CLBA should be supplemented by more appropriate placement instruments. Vandergrift and Belanger (1998) describe the background to and development of formative instruments to evaluate achievement in Canadian National Core French programmes, and argue that research shows that reactions to the instruments are positive. Both teachers and pupils regard them as beneficial for focusing and organising learning activities and find them motivating and useful for the feedback they provide to learners.

In the USA, one example of concern with school-based assessment is Manley (1995) who describes a project in a large Texas school district to develop tape-mediated tests of oral language proficiency in French, German, Spanish and Japanese, with positive outcomes.

These descriptive accounts of local and national test development contrast markedly with the literature surrounding international language proficiency examinations, like TOEFL, TWE (Test of Written English), IELTS and some Cambridge exams. Although some reports of the development of international proficiency tests are merely descriptive (for example, Charge & Taylor, 1997, and Kalter & Vossen, 1990), empirical research into various aspects of the validity and reliability of such tests is commonplace, often revealing great sophistication in analytic methodology.

This raises a continuing problem: language testing researchers tend to research and write about large-scale international tests, and not about more localised tests (including school-leaving achievement tests which are clearly relatively high-stakes). Thus, the language testing and more general educational communities lack empirical evidence about the value of many influential assessment instruments, and research often fails to address matters of educational political importance.

However, there are exceptions. For example, in connection with examination reform in Hungary, research studies have addressed issues like the use of sequencing as a test method (Alderson *et al.*, 2000b), the pairing of candidates in oral tests (Csépes *et al.*, 2000), experimentation with procedures for standard setting (Alderson, 2000a), and evidence informing ongoing debates about how many hours per week should be devoted to foreign language education in the secondary school system (Alderson, 2000b).

In commenting on the lack of international dissemination of national or regional test development work, we do not wish to deny the value of local descriptive publications. Indeed, such descriptions can serve many needs, including necessary publicity for reform work, helping teachers to understand developments, their rationale and the need for them, persuading authorities about a desired course of action or counselling against other possible actions. Publication can serve political as well as professional and academic purposes. Standard setting data can reveal what levels are achieved by the school population, including comparisons of those who started learning the language early with late-starters, those studying a first foreign language with those studying the same language as their second or third foreign language, and so on.

Language testing can inform 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 (Fekete *et al.*, 1999). What such studies have revealed has been used in in-service and pre-service teacher education and baseline studies can also be referred to in impact studies to show the effect of innovations, and to help language educators to understand how to do things more effectively. Washback studies have also been used in teacher training, both in order to influence test preparation practices, but also to encourage teachers to reflect on the reasons for their and others' practices.

LSP Testing

The development of specific purpose testing, i.e., tests in which the test content and test method are derived from a particular language use context rather than more general language use situations, can be traced back to the Temporary Registration Assessment Board (TRAB), introduced by the British General Medical Council in 1976 (see Rea-Dickins, 1987) and the development of the English Language Testing Development Unit (ELTDU) scales (Douglas, 2000). The 1980s saw the introduction of English for Academic Purposes (EAP) tests and it is these that have subsequently dominated the research and development agenda. It is important to note, however, that Language for Specific Purposes (LSP) tests are not the diametric opposite of general purpose tests. Rather, they typically fall along a continuum between general purpose tests and those for highly specialised contexts and include tests for academic purposes (e.g., the International English Language Testing System, IELTS) and for occupational or professional purposes (e.g., the Occupational English Test, OET).

Douglas (1997, 2000) identifies two aspects that typically distinguish LSP testing from general purpose testing. The first is the *authenticity* of the tasks, i.e., the test tasks share key features with the tasks that a test taker might encounter in the target language use situation. The assumption here is that the more closely the test and 'real-life' tasks are linked, the more likely it is that the test takers' performance on the test task would reflect their performance in the target situation. The second distinguishing feature of LSP testing is the *interaction between language knowledge and specific content knowledge*. This is perhaps the most crucial difference between general purpose testing and LSP testing, for in the former, any sort of background knowledge is considered to be a confounding variable that contributes construct-irrelevant variance to the test score. However, in the case of LSP testing, background knowledge constitutes an integral part of what is being tested, since it is hypothesised that test takers' language knowledge has developed within the context of their academic or professional field and that they would be disadvantaged by taking a test based on content outside that field.

The development of an LSP test typically begins with an in-depth analysis of the target language use situation, perhaps using genre analysis (see Tarone, 2001). Attention is paid to general situational features such as topics, typical lexis and grammatical structures. Specifications are then developed that take into account the specific language characteristics of the context as well as typical scenarios that occur (e.g., Plakans & Abraham, 1990; Stansfield *et al.*, 1990; Scott *et al.*, 1996; Stansfield *et al.*, 1997; Stansfield *et al.*, 2000). Particular areas of concern, quite understandably, tend to relate to issues of background knowledge and topic choice (e.g., Jensen & Hansen, 1995; Clapham, 1996; Fox *et al.*, 1997; Celestine & Cheah, 1999; Jennings *et al.*, 1999; Papajohn, 1999; Douglas, 2001a) and authenticity of task, input or, indeed, output (e.g., Lumley & Brown, 1998; Moore & Morton, 1999; Lewkowicz, 2000; Elder, 2001; Douglas, 2001a; Wu & Stansfield; 2001) and these areas of concern have been a major focus of research attention in the last decade.

Results, though somewhat mixed (cf. Jensen & Hansen, 1995 and Fox *et al.*, 1997), suggest that background knowledge and language knowledge interact differently depending on the language proficiency of the test taker. Clapham's (1996) research into subject-specific reading tests (research she conducted during and after the ELTS revision project) shows that, at least in the case of her data, the scores of neither lower nor higher proficiency test takers seemed influenced by their background knowledge. She hypothesises that for the former this was because they were most concerned with decoding the text and for the latter it was because their linguistic knowledge was sufficient for them to be able to decode the text with that alone. However, the scores of medium proficiency test takers were affected by their background knowledge. On the basis of these findings she argues that subject-specific tests are not equally valid for test takers at different levels of language proficiency.

Fox *et al.* (1997), examining the role of background knowledge in the context of the listening section of an integrated test of English for Academic Purposes (the Carleton Academic English Test, CAEL), report a slight variation on this finding. They too find a significant interaction between language proficiency and background knowledge with the scores of low proficiency test takers showing no benefit from background knowledge. However, the scores of the high proficiency candidates and analysis of their verbal protocols indicate that they did make use of their background knowledge to process the listening task.

Clapham (1996) has further shown that background knowledge is an extremely complex con-

cept. She reveals dilemmas including the difficulty of identifying with any precision the absolute specificity of an input passage and the nigh impossibility of being certain about test takers' background knowledge (particularly given that test takers often read outside their chosen academic field and might even have studied in a different academic area in the past). This is of particular concern when tests are topic-based and all the sub-tests and tasks relate to a single topic area. Jennings *et al.* (1999) and Papajohn (1999) look at the possible effect of topic, in the case of the former, for the CAEL and, in the case of the latter, in the chemistry TEACH test for international teaching assistants. They argue that the presence of topic effect would compromise the construct validity of the test whether test takers are offered a choice of topic during test administration (as with the CAEL) or not. Papajohn finds that topic does play a role in chemistry TEACH test scores and warns of the danger of assuming that subject-specificity automatically guarantees topic equivalence. Jennings *et al.* are relieved to report that choice of topic does not seem to affect test taker performance on the CAEL. However, they do note that there is a pattern in the choices made by test takers of different proficiency levels and suggest that more research is needed into the implications of these patterns for test performance.

Another particular concern of LSP test developers has been authenticity (of task, input and/or output), one example of the care taken to ensure that the test materials are authentic being Wu and Stansfield's (2001) description of the test construction procedure for the LSTE-Taiwanese (listening summary translation exam). Yet Lewkowicz (1997) somewhat puts the cat among the pigeons when she demonstrates that it is not always possible accurately to identify authentic texts from those specially constructed for testing purposes. She further problematises the valuing of authenticity in her study of a group of test takers' perceptions of an EAP test, finding that they seemed unconcerned about whether the test materials were situationally authentic or not. Indeed, they may even consider multiple-choice tests to be authentic tests of language, as opposed to tests of authentic language (Lewkowicz, 2000). (For further discussion of this topic, see Part Two of this review.)

Other test development concerns, however, are very much like those of researchers developing tests in different sub-skills. Indeed, researchers working on LSP tests have contributed a great deal to our understanding of a number of issues related to the testing of reading, writing, speaking and listening. Apart from being concerned with how best to elicit samples of language for assessment (Read, 1990), they have investigated the influence of interlocutor behaviour on test takers' performance in speaking tests (e.g., Brown & Lumley, 1997; McNamara & Lumley, 1997; Reed & Halleck, 1997). They have also studied the assumptions underpinning rating scales (Hamilton *et al.*, 1993) as well as the effect of rater variables on test scores (Brown, 1995; Lumley & McNamara, 1995) and the question of who should rate test performances – language specialists or subject specialists (Lumley, 1998).

There have also been concerns related to the interpretation of test scores. Just as in general purpose testing, LSP test developers are concerned with minimising and accounting for construct-irrelevant variables. However, this can be a particularly thorny issue in LSP testing since construct irrelevant variables can be introduced as a result of the situational authenticity of the test tasks. For instance, in his study of the chemistry TEACH test, Papajohn (1999) describes the difficulty of identifying when a teaching assistant's teaching skills (rather than language skills) are contributing to his/her test performance. He argues that test behaviours such as the provision of accessible examples or good use of the blackboard are not easily distinguished as teaching or language skills and this can result in construct-irrelevant variance being introduced into the test score. He suggests that test takers should be given specific instructions on how to present their topics, i.e., teaching tips so that teaching skills do not vary widely across performances. Stansfield *et al.* (2000) have taken a similar approach in their development of the LSTE-Taiwanese. The assessment begins with an instruction section on the summary skills needed for the test with the aim of ensuring that test performances are not unduly influenced by a lack of understanding of the task requirements.

It must be noted, however, that, because of the need for in-depth analysis of the target language use situation, LSP tests are time-consuming and expensive to produce. It is also debatable whether English for Specific Purposes (ESP) tests are more informative than a general purpose test. Furthermore, it is increasingly unclear just how 'specific' an LSP test is or can be. Indeed, more than a decade has passed since Alderson (1988) first asked the crucial question of how specific ESP testing could get. This question is recast in Elder's (2001) work on LSP tests for teachers when she asks whether for all their 'teacherliness' these tests elicit language that is essentially different from that elicited by a general language test.

An additional concern is the finding that construct relevant variables such as background knowledge and compensatory strategies interact differently with language knowledge depending on the language proficiency of the test taker (e.g., Halleck & Moder, 1995; Clapham, 1996). As a consequence of Clapham's (1996) research, the current IELTS test has no subject-specific reading texts and care is taken to ensure that the input materials are not biased for or against test takers of different disciplines. Though the extent to which this lack of bias has been achieved is debatable (see Celestine & Cheah, 1999), it can still be argued that the attempt to make texts

accessible regardless of background knowledge has resulted in the IELTS test being very weakly specific. Its claims to specificity (and indeed similar claims by many EAP tests) rest entirely on the fact that it is testing the generic language skills needed in academic contexts. This leaves it unprotected against suggestions like Clapham's (2000a) when she questions the theoretical soundness of assessing discourse knowledge that the test taker, by registering for a degree taught in English, might arguably be hoping to learn and that even a native speaker of English might lack.

Recently the British General Medical Council has abandoned its specific purpose test, the Professional and Linguistic Assessment Board (PLAB, a revised version of the TRAB), replacing it with a two-stage assessment process that includes the use of the IELTS test to assess linguistic proficiency. These developments represent the thin end of the wedge. Though the IELTS is still a specific purpose test, it is itself less so than its precursor the English Language Testing System (ELTS) and it is certainly less so than the PLAB. And so the questioning continues. Davies (2001) has joined the debate, debunking the theoretical justifications typically put forward to explain LSP testing, in particular the principle that different fields demand different language abilities. He argues that this principle is based far more on differences of content rather than on differences of language (see also Fulcher, 1999a). He also questions the view that content areas are discrete and heterogeneous.

Despite all the rumblings of discontent, Douglas (2000) stands firmly by claims made much earlier in the decade that in highly field-specific language contexts, a field-specific language test is a better predictor of performance than a general purpose test (Douglas & Selinker, 1992). He concedes that many of these contexts will be small-scale educational, professional or vocational programmes in which the number of test takers is small but maintains (Douglas, 2000: 282):

> if we want to know how well individuals can use a language in specific contexts of use, we will require a measure that takes into account both their language knowledge and their background knowledge, and their use of strategic competence in relating the salient characteristics of the target language use situation to their specific purpose language abilities. It is only by so doing ... that we can make valid interpretations of test performances.

He also suggests that the problem might not be with the LSP tests or with their specification of the target language use domain but with the assessment criteria applied. He argues (Douglas, 2001b) that just as we analyse the target language use situation in order to develop the test content and methods, we should exploit that source when we develop the assessment criteria. This might help us to avoid expecting a perfection of the test taker that is not manifested in authentic performances in the target language use situation.

But perhaps the real challenge to the field is in identifying when it is absolutely necessary to know how well someone can communicate in a specific context or if the information being sought is equally obtainable through a general-purpose language test. The answer to this challenge might not be as easily reached as is sometimes presumed.

Computer-based testing

Computer-based testing has witnessed rapid growth in the past decade and computers are now used to deliver language tests in many settings. A computer-based version of the TOEFL was introduced on a regional basis in the summer of 1998, tests are now available on CD ROM, and the Internet is increasingly used to deliver tests to users. Alderson (1996) points out that computers have much to offer language testing: not just for test delivery, but also for test construction, test compilation, response capture, test scoring, result calculation and delivery, and test analysis. They can also, of course, be used for storing tests and details of candidates.

In short, computers can be used at all stages in the test development and administration process. Most work reported in the literature, however, concerns the compilation, delivery and scoring of tests by computer. Fulcher (1999b) describes the delivery of an English language placement test over the Web and Gervais (1997) reports the mixed results of transferring a diagnostic paper-and-pencil test to the computer. Such articles set the scene for studies of computer-based testing which compare the accuracy of the computer-based test with a traditional paper-and-pencil test, addressing the advantages of a computer-delivered test in terms of accessibility and speed of results, and possible disadvantages in terms of bias against those with no computer familiarity, or with negative attitudes to computers.

This concern with bias is a recurrent theme in the literature, and it inspired a large-scale study by the Educational Testing Service (ETS), the developers of the computer-based version of the TOEFL, who needed to show that such a test would not be biased against those with no computer literacy. Jamieson *et al.* (1998) describe the development of a computer-based tutorial intended to train examinees to take the computerised TOEFL. Taylor *et al.* (1999) examine the relationship between computer familiarity and TOEFL scores, showing that those with high computer familiarity tend to score higher on the traditional TOEFL. They compare examinees with high and low computer familiarity in terms of their performance on the computer tutorial and on computerised TOEFL-like tasks. They claim that no relationship was found between computer familiarity and performance on the computerised tasks after controlling for English language proficiency. They conclude that there is no evidence of bias against candidates with low computer familiarity, but also

take comfort in the fact that all candidates will be able to take the computer tutorial before taking an operational computer-based TOEFL.

The commonest use of computers in language testing is to deliver tests adaptively (e.g., Young *et al.*, 1996). This means that the computer adjusts the items to be delivered to a candidate in the light of that candidate's success or failure on previous items. If the candidate fails a difficult item, s/he is presented with an easier item, and if s/he gets an item correct, s/he is presented with a more difficult item. This has advantages: firstly, candidates are presented with items at their level of ability, and are not faced with items that are either too easy or too difficult, and secondly, computer-adaptive tests (CATs) are typically quicker to deliver, and security is less of a problem since different candidates are presented with different items. Many authors discuss the advantages of CATs (Laurier, 1998; Brown, 1997; Chalhoub-Deville & Deville, 1999; Dunkel, 1999), but they also emphasise issues that test developers and score users must address when developing or using CATs. When designing such tests, developers have to take a number of decisions: what should the entry level be, and how is this best determined for any given population? At what point should testing cease (the so-called exit point) and what should the criteria be that determine this? How can content balance best be assured in tests where the main principle for adaptation is psychometric? What are the consequences of not allowing users to skip items, and can these consquences be ameliorated? How to ensure that some items are not presented much more frequently than others (item exposure), because of their facility, or their content? Brown and Iwashita (1996) point out that grammar items in particular will vary in difficulty according to the language background of candidates, and they show how a computer-adaptive test of Japanese resulted in very different item difficulties for speakers of English and Chinese. Thus a CAT may also need to take account of the language background of candidates when deciding which items to present, at least in grammar tests, and conceivably also in tests of vocabulary.

Chalhoub-Deville and Deville (1999) point out that, despite the apparent advantages of computer-based tests, computer-based testing relies overwhelmingly on selected response (typically multiple-choice questions) discrete-point tasks rather than performance-based items, and thus computer-based testing may be restricted to testing linguistic knowledge rather than communicative skills. However, many computer-based tests include tests of reading, which is surely a communicative skill. The question is whether computer-based testing offers any added value over paper-and-pencil reading tests: adaptivity is one possibility, although some test developers are concerned that since reading tests typically present several items on one text – what is known in the jargon as a testlet – they may not be suitable for computer-adaptivity. This concern for the inherent conservatism of computer-based testing has a long history (see Alderson, 1986a, 1986b, for example), and some claimed innovations, for example, computer-generated cloze and multiple-choice tests (Coniam, 1997, 1998) were actually implemented as early as the 1970s, and were often criticised in the literature for risking the assumption of automatic validity. But recent developments offer some hope. Burstein *et al.* (1996) argue for the relevance of new technologies in innovation in test design, construction, trialling, delivery, management, scoring, analysis and reporting. They review ways in which new input devices (e.g., voice and handwriting recognition), output devices (e.g., video, virtual reality), software such as authoring tools, and knowledge-based systems for language analysis could be used, and explore advances in the use of new technologies in computer-assisted learning materials. However, as they point out, 'innovations applied to language assessment lag behind their instructional counterparts ... the situation is created in which a relatively rich language presentation is followed by a limited productive assessment.' (1996: 245).

No doubt, this is largely due to the fact that computer-based tests require the computer to score responses. However, Burstein *et al.* (1996) argue that human-assisted scoring systems could reduce this dependency. (Human-assisted scoring systems are computer-based 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.) They also give details of free-response scoring tools which are capable of scoring responses up to 15 words long which correlate highly with human judgements (coefficients of between .89 and .98 are reported). Development of such systems for short-answer questions and for essay questions has since gone on apace. For example, ETS has developed an automated system for assessing productive language abilities, called 'e-rater'. e-rater uses natural language processing techniques to duplicate the performance of humans rating open-ended essays. Already, the system is used to rate GMAT (Graduate Management Admission Test) essays and research is ongoing for other programmes, including second/foreign language testing situations. Burstein *et al.* conclude that 'the barriers to the successful use of technology for language testing are less technical than conceptual' (1996: 253), but progress since that article was published is extremely promising.

An example of the use of IT to assess aspects of the speaking ability of second/foreign language learners of English is PhonePass. PhonePass (www.ordinate.org) is delivered over the telephone, and candidates are asked to read texts aloud, repeat heard sentences, say words opposite in meaning to heard words, and give short answers to questions. The sys-

tem uses speech recognition technology to rate responses, by comparing candidate performance to statistical models of native and non-native performance on the tasks. The system gives a score that reflects a candidate's ability to understand and respond appropriately to decontextualised spoken material, with 40% of the evaluation reflecting the fluency and pronunciation of the responses. Alderson (2000c) reports that reliability coefficients of 0.91 have been found as well as correlations with the Test of Spoken English (TSE) of 0.88 and with an ILR (Inter-agency Language Roundtable) Oral Proficiency Interview (OPI) of 0.77. An interesting feature is that the scored sample is retained on a database, classified according to the various scores assigned. This enables users to access the speech sample, in order to make their own judgements about the performance for their particular purposes, and to compare how their candidate has performed with other speech samples that have been rated either the same, or higher or lower.

In addition to e-rater and PhonePass there are a number of promising initiatives in the use of computers in testing. The listening section of the computer-based TOEFL uses photos and graphics to create context and support the content of the mini-lectures, producing stimuli that more closely approximate 'real world' situations in which people do more than just listen to voices. Moreover, candidates wear headphones, can adjust the volume control, and are allowed to control how soon the next question is presented. One innovation in test method is that candidates are required to select a visual or part of a visual; in some questions candidates must select two choices, usually out of four, and in others candidates are asked to match or order objects or texts. Moreover, candidates see and hear the test questions before the response options appear. (Interestingly, Ginther, forthcoming, suggests, however, that the use of visuals in the computer-based TOEFL listening test depresses scores somewhat, compared with traditionally delivered tests. More research is clearly needed.)

In the Reading section candidates are required to select a word, phrase, sentence or paragraph in the text itself, and other questions ask candidates to insert a sentence where it fits best. Although these techniques have been used elsewhere in paper-and-pencil tests, one advantage of their computer format is that the candidate can see the result of their choice in context, before making a final decision. Although these innovations may not seem very exciting, Bennett (1998) claims that the best way to innovate in computer-based testing is first to mount on computer what can already be done in paper-and-pencil format, with possible minor improvements allowed by the medium, in order to ensure that the basic software works well, before innovating in test method and construct. Once the delivery mechanisms work, it is argued, then computer-based deliveries can be developed that incorporate desirable innovations.

DIALANG (http://www.dialang.org) is a suite of computer-based diagnostic tests (funded by the European Union) which are available over the Internet, thus capitalising on the advantages of Internet-based delivery (see below). DIALANG uses self-assessment as an integral part of diagnosis. Users' self-ratings are combined with objective test results in order to identify a suitably difficult test for the user. DIALANG gives users feedback immediately, not only on their test scores, but also on the relationship between their test results and their self-assessment. DIALANG also gives extensive advice to users on how they can progress from their current level to the next level of language proficiency, basing this advice on the Common European Framework (Council of Europe, 2001). The interface and support language, and the language of self-assessment and of feedback, can be chosen by the test user from a list of 14 European languages. Users can decide which skill or language aspect (reading, writing, listening, grammar and vocabulary) they wish to be tested in, in any one of the same 14 European languages. Currently available test methods consist of multiple-choice, gap-filling and short-answer questions, but DIALANG has already produced CD-based demonstrations of 18 different experimental item types which could be implemented in the future, and the CD demonstrates the use of help, clue, dictionary and multiple-attempt features.

Although DIALANG is limited in its ability to assess users' productive language abilities, the experimental item types include a promising combination of self-assessment and benchmarking. Tasks for the elicitation of speaking and writing performances are administered to pilot candidates and performances are rated by human judges. Those performances on which raters achieve the greatest agreement are chosen as 'benchmarks'. A DIALANG user is presented with the same task, and, in the case of a writing task, responds via the keyboard. The user's performance is then presented on screen alongside the pre-rated benchmarks. The user can compare their own performance with the benchmarks. In addition, since the benchmarks are pre-analysed, the user can choose to see raters' comments on various features of the benchmarks, in hypertext form, and consider whether they could produce a similar quality of such features. In the case of Speaking tasks, the candidate is simply asked to imagine how they would respond to the task, rather than actually to record their performance. They are then presented with recorded benchmark performances, and are asked to estimate whether they could do better or worse than each performance. Since the performances are graded, once candidates have self-assessed themselves against a number of performances, the system can tell them roughly what level their own (imagined) performance is likely to be.

These developments illustrate some of the advantages of computer-based assessment, which make computer-based testing not only more user-friendly, but also more compatible with language pedagogy.

However, Alderson (2000c) argues the need for a research agenda, which would address the challenge of the opportunities afforded by computer-based testing and the data that can be amassed. Such an agenda would investigate the comparative advantages and added value of each form of assessment – IT-based or not IT-based. This includes issues like the effect of providing immediate feedback, support facilities, second attempts, self-assessment, confidence testing, and the like. Above all, it would seek to throw more light onto the nature of the constructs that can be tested by computer-based testing:

> What is needed above all is research that will reveal more about the validity of the tests, that will enable us to estimate the effects of the test method and delivery medium; research that will provide insights into the processes and strategies test-takers use; studies that will enable the exploration of the constructs that are being measured, or that might be measured ... And we need research into the impact of the use of the technology on learning, on learners and on the curriculum. (Alderson, 2000c: 603)

Self-assessment

The previous section has shown how computer-based testing can incorporate test takers' self-assessment of their abilities in the target language. Until the 1980s references to self-assessment were rare but since then interest in self-assessment has increased. This increase can at least in part be attributed to an increased interest in involving the learner in all phases of the learning process and in encouraging learner autonomy and decision making in (and outside) the language classroom (e.g., Blanche & Merino, 1989). The introduction of self-assessment was viewed as promising by many, especially in formative assessment contexts (Oscarson, 1989). It was considered to encourage increasing sophistication in learner awareness, helping learners to: gain confidence in their own judgement; acquire a view of evaluation that covers the whole learning process; and see errors as something helpful. It was also seen to be potentially useful to teachers, providing information on learning styles, on areas needing remediation and feedback on teaching (Barbot, 1991).

However, self-assessment also met with considerable scepticism, largely due to concerns about the ability of learners to provide accurate judgements of their achievement and proficiency. For instance, Blue (1988), while acknowledging that self-assessment is an important element in self-directed learning and that learners can play an active role in the assessment of their own language learning, argues that learners cannot self-assess unaided. Taking self-assessment data gathered from students on a pre-sessional EAP programme, he reports a poor correlation between teachers' assessments of the students and their own self-assessments. He also shows that in multicultural groups such as those typical of pre-sessional EAP courses, overestimates of language proficiency are more common than underestimates. Finally, he argues that learners' lack of familiarity with metalanguage and with the practice of discussing language proficiency in terms of its composite skills impairs their capacity for identifying their precise language learning needs.

Such concerns, however, did not dampen enthusiasm for investigations in this area and research in the 1980s was concerned with the development of self-assessment instruments and their validation (e.g., Oscarson, 1984; Lewkowicz & Moon, 1985). Consequently, a variety of approaches were developed including pupil progress cards, learning diaries, log books, rating scales and questionnaires. In the last decade the research focus has shifted towards enhancing our understanding of the evaluation techniques that were already in existence through continued validation exercises and by applying self-assessment in new contexts or in new ways.

For instance, Blanche (1990) uses standardised achievement and oral proficiency tests both for testing and for self-assessment purposes, arguing that this approach helps to circumvent the problems of training that are associated with self-assessment questionnaires. Hargan (1994) documents the use of a 'do-it-yourself' instrument for placement purposes, reporting that it results in much the same placement levels as suggested by a traditional multiple-choice test. Hargun argues that placement testing for large numbers in her context has resulted in the implementation of a traditional multiple-choice grammar-based placement test and a consequent emphasis on teaching analytic grammar skills. She believes that the 'do-it-yourself-placement' instrument might help to redress the emphasis on grammar and stem the neglect of reading and writing skills in the classroom. Carton (1993) discusses how self-assessment can become part of the learning process. He describes his use of questionnaires to encourage learners to reflect on their learning objectives and preferred modes of learning. He also presents an approach to monitoring learning that involves the learners in devising their own criteria, an approach that he argues helps learners to become more aware of their own cognitive processes.

A typical approach to validating self-assessment instruments has been to obtain concurrent validity statistics by correlating the self-assessment measure with one or more external measures of student performance (e.g., Shameem, 1998; Ross, 1998). Other approaches have included the use of multi-trait multi-method (MTMM) designs and factor analysis (Bachman & Palmer, 1989) and a split-ballot technique (Heilenman, 1990). In general, these studies have found self-assessment to be a robust method for

gathering information about learner proficiency and that the risk of cheating is low (see Barbot, 1991). However, they also indicate that some approaches to gathering self-assessment data are more effective than others. Bachman and Palmer (1989) report that learners were more able to identify what they found difficult to do in a language than what they found easy. Therefore, 'Can-do' questions were the least effective question type of the three they used in their MTMM study, while the most effective question type appeared to be that which asked about the learners' perceived difficulties with aspects of the language.

Additionally, learner experience of the self-assessment procedure and/or the language skill being assessed has been found to affect self-assessments. Heilenman (1990), in a study of the role of response effects, reports both an acquiescence effect (the tendency to respond positively to an item regardless of its content) and a tendency to overestimate ability, these tendencies being more marked among less experienced learners. Ross (1998) has found that the reliability of learners' self-assessments is affected by their experience of the skill being assessed. He suggests that when learners do not have memory of a criterion, they resort to recollections of their general proficiency in order to make their judgement. This process is more likely to be affected by the method of the self-assessment instrument and by factors such as self-flattery. He argues, therefore, for the design of instruments that are cast in terms which offer learners a reference point such as specific curricular content. In a similar finding Shameem (1998) reports that respondents' self-assessments of their oral proficiency in Fijian Hindi are less reliable at the highest levels of the self-assessment scale. Like Ross, he attributes this slip in accuracy to the respondents' lack of familiarity with the criterion measure.

Oscarson (1997) sums up progress in this area by reminding us that research in self-assessment is still relatively new. He acknowledges that conundrums remain. For instance, learner goals and interpretations need to be reconciled with external imperatives. Also self-assessment is not self-explanatory; it must be introduced slowly and learners need to be guided and supported in their use of the instruments. Furthermore, particularly when using self-assessment in multicultural groups, it is important to consider the cultural influences on self-assessment. Nevertheless, he considers the research so far to be promising. Despite residual concerns about the accuracy of self-assessment, the majority of studies report favourable results and we have already learned a great deal about the appropriate methodology to use for capturing self-assessments. However, as Oscarson points out, more work is needed, both in the study of factors that influence self-assessment ratings in various contexts and in the selection and design of materials and methods for self-assessment.

Alternative assessment

Self-assessment is one example of what is increasingly called 'alternative assessment'. 'Alternative assessment' is usually taken to mean assessment procedures which are less formal than traditional testing, which are gathered over a period of time rather than being taken at one point in time, which are usually formative rather than summative in function, are often low-stakes in terms of consequences, and are claimed to have beneficial washback effects. Although such procedures may be time-consuming and not very easy to administer and score, their claimed advantages are that they provide easily understood information, they are more integrative than traditional tests and they are more easily integrated into the classroom. McNamara (1998) makes the point that alternative assessment procedures are often developed in an attempt to make testing and assessment more responsive and accountable to individual learners, to promote learning and to enhance access and equity in education (1998: 310). Hamayan (1995) presents a detailed rationale for alternative assessment, describes different types of such assessment, and discusses procedures for setting up alternative assessment. She also provides a very useful bibliography for further reference.

A recent special issue of *Language Testing*, guest-edited by McNamara (Vol 18, 4, October 2001) reports on a symposium to discuss challenges to the current mainstream in language testing research, covering issues like assessment as social practice, democratic assessment, the use of outcomes based assessment and processes of classroom assessment. Such discussions of alternative perspectives are closely linked to so-called critical perspectives (what Shohamy calls critical language testing).

The alternative assessment movement, if it may be termed such, probably began in writing assessment, where the limitations of a one-off impromptu single writing task are apparent. Students are usually given only one, or at most two tasks, yet generalisations about writing ability across a range of genres are often made. Moreover, it is evidently the case that most writing, certainly for academic purposes but also in business settings, takes place over time, involves much planning, editing, revising and redrafting, and usually involves the integration of input from a variety of (usually written) sources. This is in clear contrast with the traditional essay which usually has a short prompt, gives students minimal input, minimal time for planning and virtually no opportunity to redraft or revise what they have produced under often stressful, time-bound circumstances. In such situations, the advocacy of portfolios of pieces of writing became a commonplace, and a whole portfolio assessment movement has developed, especially in the USA for first language writing (Hamp-Lyons & Condon, 1993, 1999) but also increasingly

for ESL writing assessment (Hamp-Lyons, 1996) and also for the assessment of foreign languages (French, Spanish, German, etc.) writing assessment.

Although portfolio assessment in other subject areas (art, graphic design, architecture, music) is not new, in foreign language education portfolios have been hailed as a major innovation, supposedly overcoming the drawbacks of traditional assessment. A typical example is Padilla *et al.* (1996) who describe the design and implementation of portfolio assessment in Japanese, Chinese, Korean and Russian, to assess growth in foreign language proficiency. They make a number of practical recommendations to assist teachers wishing to use portfolios in progress assessment.

Hughes Wilhelm (1996) describes how portfolio assessment was integrated with criterion-referenced grading in a pre-university English for academic purposes programme, together with the use of contract grading and collaborative revision of grading criteria. It is claimed that such an assessment scheme encourages learner control whilst maintaining standards of performance.

Short (1993) discusses the need for better assessment models for instruction where content and language instruction are integrated. She describes examples of the implementation of a number of alternative assessment measures, such as checklists, portfolios, interviews and performance-tasks, in elementary and secondary school integrated content and language classes.

Alderson (2000d) describes a number of alternative procedures for assessing reading, including checklists, teacher-pupil conferences, learner diaries and journals, informal reading inventories, classroom reading aloud sessions, portfolios of books read, self-assessments of progress in reading, and the like.

Many of the accounts of alternative assessment are for classroom-based assessment, often for assessing progress through a programme of instruction. Gimenez (1996) gives an account of the use of process assessment in an ESP course; Bruton (1991) describes the use of continuous assessment over a full school year in Spain, to measure achievement of objectives and learner progress. Haggstrom (1994) describes ways she has successfully used a video camera and task-based activities to make classroom-based oral testing more communicative and realistic, less time-consuming for the teacher, and more enjoyable and less stressful for students. Lynch (1988) describes an experimental system of peer evaluation using questionnaires in a pre-sessional EAP summer programme, to assess speaking abilities. He concludes that this form of evaluation had a marked effect on the extent to which speakers took their audience into account. Lee (1989) discusses how assessment can be integrated with the learning process, illustrating her argument with an example where pupils prepare, practise and perform a set task in Spanish together. She offers practical tips for how teachers can reduce the amount of paperwork involved in classroom assessment of this sort. Sciarone (1995) discusses the difficulties of monitoring learning with large groups of students (in contrast with that of individuals) and describes the use, with 200 learners of Dutch, of a simple monitoring tool (a personal computer) to keep track of the performance of individual learners on a variety of learning tasks.

Typical of these accounts, however, is the fact that they are descriptive and persuasive, rather than research-based, or empirical studies of the advantages and disadvantages of 'alternative assessment'. Brown and Hudson (1998) present a critical overview of such approaches, criticising the evangelical way in which advocates assert the value and indeed validity of their procedures without any evidence to support their assertions. They point out that there is no such thing as automatic validity, a claim all too often made by the advocates of alternative assessment. Instead of 'alternative assessment', they propose the term 'alternatives in assessment', pointing out that there are many different testing methods available for assessing student learning and achievement. They present a description of these methods, including selected-response techniques, constructed-response techniques and personal-response techniques. Portfolio and other forms of 'alternative assessment' are classified under the latter category, but Brown and Hudson emphasise that they should be subject to the same criteria of reliability, validity and practicality as any other assessment procedure, and should be critically evaluated for their 'fitness for purpose', what Bachman and Palmer (1996) called 'usefulness'. Hamp-Lyons (1996) concludes that portfolio scoring is less reliable than traditional writing rating; little training is given and raters may be judging the writer as much as the writing. Brown and Hudson emphasise that decisions for use of any assessment procedure should be informed by considerations of consequences (washback), the significance and need for, and value of, feedback based on the assessment results, and the importance of using multiple sources of information when making decisions based on assessment information.

Clapham (2000b) makes the point that many alternative assessment procedures are not pre-tested and trialled, their tasks and mark schemes are therefore of unknown or even dubious quality, and despite face validity, they may not tell the user very much at all about learners' abilities.

In short, as Hamayan (1995) admits, alternative assessment procedures have yet to 'come of age', not only in terms of demonstrating beyond doubt their usefulness, in Bachman and Palmer's terms, but also in terms of being implemented in mainstream assessment, rather than in informal class-based assessment. She argues that consistency in the application of alternative assessment is still a problem, that mech-

anisms for thorough self-criticism and evaluation of alternative assessment procedures are lacking, that some degree of standardisation of such procedures will be needed if they are to be used for high-stakes assessment, and that the financial and logistic viability of such procedures remains to be demonstrated.

Assessing young learners

Finally, in this first part of our review, we consider recent developments in the assessment of young learners, an area where it is often argued that alternative assessment procedures are more appropriate than formal testing procedures. Typically considered to apply to the assessment of children between the ages of 5 and 12 (but also including much younger and slightly older children), the assessment of young learners dates back to the 1960s. However, research interest in this area is relatively new and the last decade has witnessed a plethora of studies (e.g., Low *et al.*, 1993; McKay *et al.*, 1994; Edelenbos & Johnstone, 1996; Breen *et al.*, 1997; Leung & Teasdale, 1997; TESOL, 1998; Blondin *et al.*, 1998). This trend can be largely attributed to three factors. Firstly, second language teaching (particularly English) to children in the pre-primary and primary age groups both within mainstream education and by commercial organisations, has mushroomed. Secondly, it is recognised that classrooms have become increasingly multi-cultural and, particularly in the context of Australia, Canada, the United States and the UK, many learners are speakers of English as an additional/second language (rather than heritage speakers of English). Thirdly, the decade has seen an increased proliferation, within mainstream education, of teaching and learning standards (e.g., the National Curriculum Guidelines in England and Wales) and demands for accountability to stakeholders.

The research that has resulted falls broadly into three areas: the assessment of language delay and/or impairment, the assessment of young learners with English as an additional/second language, and the assessment of foreign languages in primary/elementary school.

Changes in the measurement of language delay and/or impairment have been attributed to theoretical and practical advances in speech and language therapy. It is claimed that these advances have, in turn, wrought changes in the scope of what is involved in language assessment and in the methods by which it takes place (Howard *et al.*, 1995). Resulting research has included reflection on the predictive validity of tests involving language production that are used as standard screening for language delay in children as young as 18 months (particularly in the light of research evidence that production and comprehension are not functionally discrete before 28 months) (Boyle *et al.*, 1996). Other research, however, has looked at the nature of the language disorder. Windsor (1999) investigates the effect of semantic inconsistency on sentence grammaticality judgements for children with and without language-learning disabilities (LD), finding that children with LD differed most from their chronological age-group peers in the identification of ungrammatical sentences and that it is important to consider the effect on performance of competing linguistic information in the task. Holm *et al.* (1999) have developed a phonological assessment procedure for bilingual children, using this assessment to describe the phonological development, in each language, of normally developing bilingual children as well as of two bilingual children with speech disorders. They conclude that the normal phonological development of bilingual children differs from monolingual development in each of the languages and that the phonological output of bilingual children with speech disorders reflects a single underlying deficit. The findings of these studies have implications for the design of assessment tools as well as for the need to identify appropriate norms against which to measure performance on the assessments.

Such issues, particularly the identification of appropriate norms of performance, are also important in studies of young learners' readiness to access mainstream education in a language other than their heritage language. Recent research involving learners of English as an additional or second language (EAL/ESL) has benefited from work in the 1980s (e.g., Stansfield, 1981; Cummins, 1984a, 1984b; Barrs *et al.*, 1988; Trueba, 1989) which problematised the use of standardised tests that had been normed on monolingual learners of English. The equity considerations they raised, particularly the *false positive* diagnosis of EAL/ESL learners as having learning disabilities, has resulted in the development of EAL/ESL learner 'profiles' (also called standards/benchmarks/scales) (see NLLIA, 1993; Australian Education Council, 1994; TESOL, 1998). Research has also focused on the provision of guidance for teachers when monitoring and reporting on learner progress (see McKay & Scarino, 1991; Genesee & Hamayan, 1994; Law & Eckes, 1995). Curriculum-based age-level tasks have also been developed to help teachers observe performance and place learners on a common framework/standard (Lumley *et al.*, 1993).

However, these directions, though productive, have not been unproblematic, not least because they imply (and indeed encourage) differential assessment for EAL/ESL learners in order for individual students' needs to be identified and addressed. This can result in tension between the concerns of the educational system for ease of administration, appearances of equity and accountability and those of teachers for support in teaching and learning (see Brindley, 1995). Indeed, Australia and England and Wales have now introduced standardised testing

for all learners regardless of language background. The latter two countries are purportedly following a policy of entitlement for all but, as McKay (2000) argues, their motives are far more likely to be to simplify/rationalise reporting in order to make comparisons across schools and on which to predicate funding. Furthermore, and somewhat paradoxically, as Leung and Teasdale (1996) have established, the use of standardised attainment targets does not result in more equitable treatment of learners, because teachers implicitly apply native-speaker norms in making judgements of EAL/ESL learner performances.

Latterly, research has focused on classroom-based teacher assessment, looking, in the case of Rea-Dickins and Gardner (2000), at the constructs underlying formative and summative assessment and, in the case of Teasdale and Leung (2000), at the epistemic and practical challenges for alternative assessment. The overriding conclusion of both studies is that 'insufficient research has been done to establish what, if any, elements of assessment for learning and assessment as measurement are compatible' (Teasdale & Leung, 2000: 180), a concern no doubt shared by researchers studying the introduction of assessment of foreign languages in primary/elementary schools.

Indeed, the growing tendency to introduce a foreign language at the primary school level has resulted in a parallel growth in interest in how this early learning might be assessed. This research focuses on both formative (e.g., Hasselgren, 1998; Gattullo, 2000; Hasselgren, 2000; Zangl, 2000) and summative assessment (Johnstone, 2000; Edelenbos & Vinje, 2000) and is primarily concerned with how young learners' foreign language skills might be assessed, with an emphasis on identifying what learners can do. Motivated in many cases by a need to evaluate the effectiveness of language programmes (e.g., Carpenter *et al.*, 1995; Edelenbos & Vinje, 2000), these studies document the challenges of designing tests for young learners. In doing so they cite, among other factors: the learners' need for fantasy and fun, the potentially detrimental effect of perceived 'failure' on future language learning, the need to design tasks that are developmentally appropriate and comparable for children of different language abilities who have studied in different schools/language programmes and the potential problem inherent in tasks which encourage children to interact with an unfamiliar adult in the test situation (see Carpenter *et al.*, 1995; Hasselgren, 1998, 2000). The studies also reflect a desire to understand how teachers implement assessment (Gatullo, 2000) as well as a need for inducting teachers into assessment practices in contexts where there is no tradition of assessment (Hasselgren, 1998).

Recent years have also seen a phenomenal increase in the number of commercial language classes for young learners with a consequent market for certification of progress. The latest additions to the certificates available are the Saxoncourt Tests for Young Learners of English (STYLE) (http://www.saxoncourt.com/publishing.htm) and a suite of tests for young learners developed by the University of Cambridge Local Examinations Syndicate (UCLES): Starters, Movers and Flyers (http://www.cambridge-efl.org/exam/young/bg_yle.htm)

In the development of the latter, the cognitive development of young learners has purportedly been taken into account and though certificates are issued, these are intended to reward young learners for what they can do. By adopting this approach it is hoped that the tests will be used to find out what the learners already know/have learned and to check if teaching objectives have been achieved (Wilson, 2001).

It is clear that, despite an avowed preference for teacher-based formative assessment, recent research on assessing young learners documents a growth in formal assessment and ongoing research exemplifies the movement towards greater standardisation of assessment activities and measures of attainment. Furthermore, the expansion in formal assessment has led to increased specification of the language targets young learners might plausibly be expected to reach and indicates the spread of centrally specified curriculum goals. It seems that the field has moved forward in its understanding of the assessment needs of young learners yet has been pressed back by economic considerations. The challenge in the next decade will perhaps lie in addressing the tension between these competing agendas.

In this first part of the two-part review of language testing and assessment, we have reviewed relatively new concerns in language testing, beginning with an account of research into washback, and then moving on to discuss issues in the ethics and politics of language testing and the development of standards for language tests. After describing trends in testing on a national level and developments in testing for specific purposes, we surveyed developments in computer-based testing before discussing self-assessment and alternative assessment. Finally we reviewed the assessment of young learners.

In the second part of this review, to appear in April 2002, we describe developments in what are basically rather traditional concerns in language testing research, looking at the major language constructs (reading, listening, and so on) but in the context of a new approach to validity and validation, sometimes known as the Messick approach, or construct validation.

References

ALDERSON, J. C. (1986a). Computers in language testing. In G. N. Leech & C. N. Candlin (Eds.), *Computers in English language education and research* (pp. 99–111). London: Longman.

ALDERSON, J. C. (1986b). Innovations in language testing? In M.

Portal (Ed.), *Innovations in language testing* (pp. 93–105). Windsor: NFER/Nelson.

ALDERSON, J. C. (1988). Testing English for Specific Purposes: how specific can we get? *ELT Documents, 127*, 16–28.

ALDERSON, J. C. (1991). Language testing in the 1990s: How far have we got? How much further have we to go? In S. Anivan (Ed.), *Current developments in language testing* (Vol. 25, pp. 1–26). Singapore: SEAMEO Regional Language Centre.

ALDERSON, J. C. (1996). Do corpora have a role in language assessment? In J. Thomas & M. Short (Eds.), *Using corpora for language research* (pp. 248–59). Harlow: Longman.

ALDERSON, J. C. (1997). Ethics and language testing. *Paper presented at the annual TESOL Convention, Orlando, Florida.*

ALDERSON, J. C. (1998). Testing and teaching: the dream and the reality. *novELTy, 5*(4), 23–37.

ALDERSON, J. C. (1999). *What does PESTI have to do with us testers?* Paper presented at the International Language Education Conference, Hong Kong.

ALDERSON, J. C. (2000a). Levels of performance. In J. C. Alderson, E. Nagy, & E. Öveges (Eds.), *English language education in Hungary, Part II: Examining Hungarian learners' achievements in English.* Budapest: The British Council.

ALDERSON, J. C. (2000b). Exploding myths: Does the number of hours per week matter? *novELTy, 7*(1), 17–32.

ALDERSON, J. C. (2000c). Technology in testing: the present and the future. *System 28* (4) 593–603.

ALDERSON, J. C. (2000d). *Assessing reading.* Cambridge: Cambridge University Press.

ALDERSON, J. C. (2001a). Testing is too important to be left to the tester. *Paper presented at the 3rd Annual Language Testing Symposium, Dubai, United Arab Emirates.*

ALDERSON, J. C. (2001b). *The lift is being fixed. You will be unbearable today (Or why we hope that there will not be translation on the new English érettsegi).* Paper presented at the Magyar Macmillan Conference, Budapest, Hungary.

ALDERSON, J. C. & BUCK, G. (1993). Standards in testing: a study of the practice of UK examination boards in EFL/ESL testing. *Language Testing, 10*(1), 1–26.

ALDERSON, J. C., CLAPHAM, C. & WALL, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

ALDERSON, J. C. & HAMP-LYONS, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing, 13*(3), 280–97.

ALDERSON, J. C., NAGY, E. & ÖVEGES, E. (Eds.) (2000a). *English language education in Hungary, Part II: Examining Hungarian learners' achievements in English.* Budapest: The British Council.

ALDERSON, J. C., PERCSICH, R. & SZABO, G. (2000b). Sequencing as an item type. *Language Testing, 17* (4), 423–47.

ALDERSON, J. C. & WALL, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–29.

ALTE (1998). *ALTE handbook of European examinations and examination systems.* Cambridge: UCLES.

AUSTRALIAN EDUCATION COUNCIL (1994). *ESL Scales.* Melbourne: Curriculum Corporation.

BACHMAN, L. F. & PALMER, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*(1), 14–29.

BACHMAN, L. F. & PALMER, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

BAILEY, K. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257–79.

BAKER, C. (1988). Normative testing and bilingual populations. *Journal of Multilingual and Multicultural Development, 9*(5), 399–409.

BANERJEE, J., CLAPHAM, C., CLAPHAM, P. & WALL, D. (Eds.) (1999). *ILTA language testing bibliography 1990–1999, First edition.* Lancaster, UK: Language Testing Update.

BARBOT, M.-J. (1991). New approaches to evaluation in self-access learning (trans. from French). *Études de Linguistique Appliquée, 79*, 77–94.

BARNES, A., HUNT, M. & POWELL, B. (1999). Dictionary use in the teaching and examining of MFLs at GCSE. *Language Learning Journal, 19*, 19–27.

BARNES, A. & POMFRETT, G. (1998). Assessment in German at KS3: how can it be consistent, fair and appropriate? *Deutsch: Lehren und Lernen, 17*, 2–6.

BARRS, M., ELLIS, S., HESTER, H. & THOMAS, A. (1988). *The Primary Language Record: A handbook for teachers.* London: Centre for Language in Primary Education.

BENNETT, R. E. (1998). *Reinventing assessment: speculations on the future of large-scale educational testing.* Princeton, New Jersey: Educational Testing Service.

BHGEL, K. & LEIJN, M. (1999). New exams in secondary education, new question types. An investigation into the reliability of the evaluation of open-ended questions in foreign-language exams. *Levende Talen, 537*, 173–81.

BLANCHE, P. (1990). Using standardised achievement and oral proficiency tests for self-assessment purposes: the DLIFLC study. *Language Testing, 7*(2), 202–29.

BLANCHE, P. & MERINO, B. J. (1989). Self-assessment of foreign language skills: implications for teachers and researchers. *Language Learning, 39*(3), 313–40.

BLONDIN, C., CANDELIER, M., EDELENBOS, P., JOHNSTONE, R., KUBANEK-GERMAN, A. & TAESCHNER, T. (1998). *Foreign languages in primary and pre-school education: context and outcomes. A review of recent research within the European Union.* London: CILT.

BLUE, G. M. (1988). Self assessment: the limits of learner independence. *ELT Documents, 131*, 100–18.

BOLOGNA DECLARATION (1999) Joint declaration of the European Ministers of Education convened in Bologna on the 19th of June 1999. http://europa.eu.int/comm/education/socrates/erasmus/bologna.pdf

BOYLE, J., GILLHAM, B. & SMITH, N. (1996). Screening for early language delay in the 18–36 month age-range: the predictive validity of tests of production and implications for practice. *Child Language Teaching and Therapy, 12*(2), 113–27.

BREEN, M. P., BARRATT-PUGH, C., DEREWIANKA, B., HOUSE, H., HUDSON, C., LUMLEY, T., & ROHL, M. (Eds.) (1997). *Profiling ESL children: how teachers interpret and use national and state assessment frameworks* (Vol. 1). Commonwealth of Australia: Department of Employment, Education, Training and Youth Affairs.

BRINDLEY, G. (1995). *Assessment and reporting in language learning programs: Purposes, problems and pitfalls.* Plenary presentation at the International Conference on Testing and Evaluation in Second Language Education, Hong Kong University of Science and Technology, 21–24 June 1995.

BRINDLEY, A. (1998). Outcomes-based assessment and reporting in language learning programmes: a review of the issues. *Language Testing, 15*(1), 45–85.

BRINDLEY, G. (2001). Outcomes-based assessment in practice: some examples and emerging insights. *Language Testing, 18*(4), 393–407.

BROWN, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1–15.

BROWN, A. & IWASHITA, N. (1996). Language background and item difficulty: the development of a computer-adaptive test of Japanese. *System, 24*(2), 199–206.

BROWN, A. & LUMLEY, T. (1997). Interviewer variability in specific-purpose language performance tests. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (137–50). Jyväskylä: Centre for Applied Language Studies, University of Jyväskylä.

BROWN, J. D. (1997). Computers in language testing: present research and some future directions. *Language Learning and Technology, 1*(1), 44–59.

BROWN, J. D. & HUDSON, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653–75.

BRUTON, A. (1991). Continuous assessment in Spanish state schools. *Language Testing Update, 10*, 14–20.

BUCKBY, M. (1999). The use of the target language at GCSE. *Language Learning Journal, 19*, 4–11.

BURSTEIN, J., FRASE, L. T., GINTHER, A. & GRANT, L. (1996). Technologies for language assessment. *Annual Review of Applied Linguistics, 16*, 240–60.

CARPENTER, K., FUJII, N. & KATAOKA, H. (1995). An oral interview procedure for assessing second language abilities in children. *Language Testing, 12*(2), 157–81.

CARROLL, B. J. & WEST, R. (1989). *ESU Framework: Performance scales for English language examinations*. Harlow: Longman.

CARTON, F. (1993). Self-evaluation at the heart of learning. *Le Français dans le Monde* (special number), 28–35.

CELESTINE, C. & CHEAH, S. M. (1999). The effect of background disciplines on IELTS scores. In R. Tulloh (Ed.), *IELTS Research Reports 1999* (Vol. 2, 36–51). Canberra: IELTS Australia Pty Limited.

CHALHOUB-DEVILLE, M. & DEVILLE, C. (1999). Computer-adaptive testing in second language contexts. *Annual Review of Applied Linguistics, 19*, 273–99.

CHAMBERS, F. & RICHARDS, B. (1992). Criteria for oral assessment. *Language Learning Journal, 6*, 5–9.

CHAPELLE, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254–72.

CHARGE, N. & TAYLOR, L. B. (1997). Recent developments in IELTS. *ELT Journal, 51*(4), 374–80.

CHEN, Z. & HENNING, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155–63.

CHENG, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education 11*(1), 38–54.

CLAPHAM, C. (1996). *The development of IELTS: a study of the effect of background knowledge on reading comprehension* (Studies in Language Testing Series, Vol. 4). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

CLAPHAM, C. (2000a). Assessment for academic purposes: where next? *System, 28*, 511–21.

CLAPHAM, C. (2000b). Assessment and testing. *Annual Review of Applied Linguistics, 20*, 147–61.

CONIAM, D. (1994). Designing an ability scale for English across the range of secondary school forms. *Hong Kong Papers in Linguistics and Language Teaching, 17*, 55–61.

CONIAM, D. (1995). Towards a common ability scale for Hong Kong English secondary-school forms. *Language Testing, 12*(2), 182–93.

CONIAM, D. (1997). A computerised English language proofing cloze program. *Computer-Assisted Language Learning, 10*(1), 83–97.

CONIAM, D. (1998). From text to test, automatically – an evaluation of a computer cloze-test generator. *Hong Kong Journal of Applied Linguistics, 3*(1), 41–60.

COUNCIL OF EUROPE (2001). *A Common European Framework of reference for learning, teaching and assessment*. Cambridge: Cambridge University Press.

CSÉPES, I., SULYOK, A. & ÖVEGES, E. (2000). The pilot speaking examinations. In J. C. Alderson, E. Nagy & E. Öveges (Eds.), *English language education in Hungary, Part II: Examining Hungarian learners' achievements in English*. Budapest: The British Council.

CUMMING, A. (1994). Does language assessment facilitate recent immigrants' participation in Canadian society? *TESL Canada Journal, 11*(2), 117–33.

CUMMING, A. (1995). Changing definitions of language proficiency: functions of language assessment in educational programmes for recent immigrant learners of English in Canada. *Journal of the CAAL, 17*(1), 35–48.

CUMMINS, J. (1984a). *Bilingualism and special education: Issues in assessment and pedagogy*. Clevedon, England: Multilingual Matters.

CUMMINS, J. (1984b). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (Ed.), *Language proficiency and academic achievement* (Vol. 10). Clevedon, England: Multilingual Matters.

DAVIDSON, F. (1994). Norms appropriacy of achievement tests: Spanish-speaking children and English children's norms. *Language Testing, 11*(1), 83–95.

DAVIES, A. (1978). Language testing: survey articles 1 and 2. *Language Teaching and Linguistics Abstracts, 11*, 145–59 and 215–31.

DAVIES, A. (1997). Demands of being professional in language testing. *Language Testing, 14*(3), 328–39.

DAVIES, A. (2001). The logic of testing Languages for Specific Purposes. *Language Testing, 18*(2), 133–47.

DE JONG, J. H. A. L. (1992). Assessment of language proficiency in the perspective of the 21st century. *AILA Review, 9*, 39–45.

DOLLERUP, C., GLAHN, E. & ROSENBERG HANSEN, C. (1994). 'Sprogtest': a smart test (or how to develop a reliable and anonymous EFL reading test). *Language Testing, 11*(1), 65–81.

DOUGLAS, D. (1995). Developments in language testing. *Annual Review of Applied Linguistics, 15*, 167–87.

DOUGLAS, D. (1997). Language for specific purposes testing. In C. Clapham & D. Corson (Eds.), *Language testing and assessment* (Vol. 7, 111–19). Dordrecht, The Netherlands: Kluwer Academic Publishers.

DOUGLAS, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

DOUGLAS, D. (2001a). Three problems in testing language for specific purposes: authenticity, specificity and inseparability. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. F. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: essays in honour of Alan Davies* (Studies in Language Testing Series, Vol. 11, 45–51). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

DOUGLAS, D. (2001b). Language for Specific Purposes assessment criteria: where do they come from? *Language Testing, 18*(2), 171–85.

DOUGLAS, D. & SELINKER, L. (1992). Analysing oral proficiency test performance in general and specific-purpose contexts. *System, 20*(3), 317–28.

DUNKEL, P. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning and Technology, 2*(2), 77–93.

EDELENBOS, P. & JOHNSTONE, R. (Eds.). (1996). *Researching languages at primary school: some European perspectives*. London: CILT, in collaboration with Scottish CILT and GION.

EDELENBOS, P. & VINJE, M. P. (2000). The assessment of a foreign language at the end of primary (elementary) education. *Language Testing, 17*(2), 144–62.

ELDER, C. (1997). What does test bias have to do with fairness? *Language Testing, 14*(3), 261–77.

ELDER, C. (2001). Assessing the language proficiency of teachers: are there any border controls? *Language Testing, 18*(2), 149–70.

FEKETE, H., MAJOR, E. & NIKOLOV, M. (Eds.) (1999). *English language education in Hungary: A baseline study*. Budapest: The British Council.

FOX, J., PYCHYL, T. & ZUMBO, B. (1997). An investigation of background knowledge in the assessment of language proficiency. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (367–83). Jyväskylä: University of Jyväskylä.

FULCHER, G. (1999a). Assessment in English for Academic Purposes: putting content validity in its place. *Applied Linguistics, 20*(2), 221–36.

FULCHER, G. (1999b). Computerising an English language placement test. *ELT Journal, 53*(4), 289–99.

FULCHER, G. & BAMFORD, R. (1996). I didn't get the grade I need. Where's my solicitor? *System, 24*(4), 437–48.

GATTULLO, F. (2000). Formative assessment in ELT primary

(elementary) classrooms: an Italian case study. *Language Testing, 17*(2), 278–88.

GENESEE, F. & HAMAYAN, E. V. (1994). Classroom-based assessment. In F. Genesee (Ed.), *Educating second language children.* Cambridge: Cambridge University Press.

GERVAIS, C. (1997). Computers and language testing: a harmonious relationship? *Francophonie, 16*, 3–7.

GIMENEZ, J. C. (1996). Process assessment in ESP: input, throughput and output. *English for Specific Purposes, 15*(3), 233–41.

GINTHER, A. (forthcoming). Context and content visuals and performance on listening comprehension stimuli. *Language Testing.*

GROOT, P. J. M. (1990). Language testing in research and education: the need for standards. *AILA Review, 7*, 9–23.

GUILLON, M. (1997). L'évaluation ministerielle en classe de seconde en anglais. *Les Langues Modernes, 2*, 32–39.

HAGGSTROM, M. (1994). Using a videocamera and task-based activities to make classroom oral testing a more realistic communicative experience. *Foreign Language Annals, 27*(2), 161–75.

HAHN, S., STASSEN, T. & DESCHKE, C. (1989). Grading classroom oral activities: effects on motivation and proficiency. *Foreign Language Annals, 22*(3), 241–52.

HALLECK, G. B. & MODER, C. L. (1995). Testing language and teaching skills of international teaching assistants: the limits of compensatory strategies. *TESOL Quarterly, 29*(4), 733–57.

HAMAYAN, E. (1995). Approaches to alternative assessment. *Annual Review of Applied Linguistics, 15*, 212–26.

HAMILTON, J., LOPES, M., MCNAMARA, T. & SHERIDAN, E. (1993). Rating scales and native speaker performance on a communicatively oriented EAP test. *Language Testing, 10*(3), 337–53.

HAMP-LYONS, L. (1996). Applying ethical standards to portfolio assessment of writing in English as a second language. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (Studies in Language Testing Series, Vol. 3, 151–64). Cambridge: Cambridge University Press.

HAMP-LYONS, L. (1997). Washback, impact and validity: ethical concerns. *Language Testing, 14*(3), 295–303.

HAMP-LYONS, L. (1998). Ethics in language testing. In C. M. Clapham & D. Corson (Eds.), *Language testing and assessment* (Vol. 7). Dordrecht, The Netherlands: Kluwer Academic Publishing.

HAMP-LYONS, L. & CONDON, W. (1993). Questioning assumptions about portfolio-based assessment. *College Composition and Communication, 44*(2), 176–90.

HAMP-LYONS, L. & CONDON, W. (1999). *Assessing college writing portfolios: principles for practice, theory, research.* Cresskill, NJ: Hampton Press.

HARGAN, N. (1994). Learner autonomy by remote control. *System, 22*(4), 455–62.

HASSELGREN, A. (1998). *Small words and good testing.* Unpublished PhD dissertation, University of Bergen, Bergen.

HASSELGREN, A. (2000). The assessment of the English ability of young learners in Norwegian schools: an innovative approach. *Language Testing, 17*(2), 261–77.

HAWTHORNE, L. (1997). The political dimension of language testing in Australia. *Language Testing, 14*(3), 248–60.

HEILENMAN, L. K. (1990). Self-assessment of second language ability: the role of response effects. *Language Testing, 7*(2), 174–201.

HENRICHSEN, L. E. (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan, 1956–1968.* New York: Greenwood Press.

HOLM, A., DODD, B., STOW, C. & PERT, S. (1999). Identification and differential diagnosis of phonological disorder in bilingual children. *Language Testing, 16*(3), 271–92.

HOWARD, S., HARTLEY, J. & MUELLER, D. (1995). The changing face of child language assessment: 1985–1995. *Child Language Teaching and Therapy, 11*(1), 7–22.

HUGHES, A. (1993). Backwash and TOEFL 2000. Unpublished manuscript, University of Reading.

HUGHES WILHELM, K. (1996). Combined assessment model for EAP writing workshop: portfolio decision-making, criterion-referenced grading and contract negotiation. *TESL Canada Journal, 14*(1), 21–33.

HURMAN, J. (1990). Deficiency and development. *Francophonie, 1*, 8–12.

ILTA – INTERNATIONAL LANGUAGE TESTING ASSOCIATION (1997). Code of practice for foreign/ second language testing. Lancaster: ILTA. [Draft, March, 1997].

ILTA – INTERNATIONAL LANGUAGE TESTING ASSOCIATION. Code of Ethics. [http://www.surrey.ac.uk/ELI/ltrfile/ltr-frame.html]

JAMIESON, J., TAYLOR, C., KIRSCH, I. & EIGNOR, D. (1998). Design and evaluation of a computer-based TOEFL tutorial. *System, 26*(4), 485–513.

JANSEN, H. & PEER, C. (1999). Using dictionaries with national foreign-language examinations for reading comprehension. *Levende Talen, 544*, 639–41.

JENNINGS, M., FOX, J., GRAVES, B. & SHOHAMY, E. (1999). The test-takers' choice: an investigation of the effect of topic on language-test performance. *Language Testing, 16*(4), 426–56.

JENSEN, C. & HANSEN, C. (1995) The effect of prior knowledge on EAP listening-test performance, *Language Testing, 12*(1), 99–119.

JOHNSTONE, R. (2000). Context-sensitive assessment of modern languages in primary (elementary) and early secondary education: Scotland and the European experience. *Language Testing, 17*(2), 123–43.

KALTER, A. O. & VOSSEN, P. W. J. E. (1990). EUROCERT: an international standard for certification of language proficiency. *AILA Review, 7*, 91–106.

KHANIYAH, T. R. (1990a). *Examinations as instruments for educational change: Investigating the washback effect of the Nepalese English exams.* Unpublished PhD dissertation, University of Edinburgh, Edinburgh.

KHANIYAH, T. R. (1990b). The washback effect of a textbook-based test. *Edinburgh Working Papers in Applied Linguistics, 1*, 48–58.

KIEWEG, W. (1992). Leistungsmessung im Fach Englisch: Praktische Vorschläge zur Konzeption von Lernzielkontrollen. *Fremdsprachenunterricht, 45*(6), 321–32.

KIEWEG, W. (1999). Allgemeine Gütekriterien für Lernzielkontrollen (Common standards for the control of learning). *Der Fremdsprachliche Unterricht Englisch, 37*(1), 4–11.

LAURIER, M. (1998). Méthodologie d'évaluation dans des contextes d'apprentissage des langages assistés par des environnements informatiques multimédias. *Études de Linguistique Appliquée, 110*, 247–55.

LAW, B. & ECKES, M. (1995). *Assessment and ESL.* Winnipeg, Canada: Peguis.

LEE, B. (1989). Classroom-based assessment – why and how? *British Journal of Language Teaching, 27*(2), 73–6.

LEUNG, C. & TEASDALE, A. (1996). English as an additional language within the National Curriculum: A study of assessment practices. *Prospect, 12*(2), 58–68.

LEUNG, C. & TEASDALE, A. (1997). What do teachers mean by speaking and listening: a contextualised study of assessment in the English National Curriculum. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *New contexts, goals and alternatives in language assessment* (291–324). Jyväskylä: University of Jyväskylä.

LEWKOWICZ, J. A. (1997). *Investigating authenticity in language testing.* Unpublished PhD dissertation, Lancaster University, Lancaster.

LEWKOWICZ, J. A. (2000). Authenticity in language testing: some outstanding questions. *Language Testing, 17*(1), 43–64.

LEWKOWICZ, J. A., & MOON, J. (1985). Evaluation, a way of involving the learner. In J. C. Alderson (Ed.), *Lancaster Practical*

Papers in English Language Education (Vol. 6: Evaluation), 45–80. Oxford: Pergamon Press.

LI, K. C. (1997). The labyrinth of exit standard controls. *Hong Kong Journal of Applied Linguistics, 2*(1), 23–38.

LIDDICOAT, A. (1996). The Language Profile: oral interaction. *Babel, 31*(2), 4–7, 35.

LIDDICOAT, A. J. (1998). Trialling the languages profile in the A.C.T. *Babel, 33*(2), 14–38.

LOW, L., DUFFIELD, J., BROWN, S. & JOHNSTONE, R. (1993). *Evaluating foreign languages in Scottish primary schools: report to Scottish Office.* Stirling: University of Stirling: Scottish CILT.

LUMLEY, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes, 17*(4), 347–67.

LUMLEY, T. & BROWN, A. (1998). Authenticity of discourse in a specific purpose test. In E. Li & G. James (Eds.), *Testing and evaluation in second language education* (22–33). Hong Kong: The Language Centre, The University of Science and Technology.

LUMLEY, T. & MCNAMARA, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*(1), 54–71.

LUMLEY, T., RASO, E. & MINCHAM, L. (1993). Exemplar assessment activities. In NLLIA (Ed.), *NLLIA ESL Development: Language and Literacy in Schools.* Canberra: National Languages and Literacy Institute of Australia.

LYNCH, B. (1997). In search of the ethical test. *Language Testing, 14*(3), 315–27.

LYNCH, B. & DAVIDSON, F. (1994). Criterion-referenced test development: linking curricula, teachers and tests. *TESOL Quarterly, 28*(4), 727–43.

LYNCH, T. (1988). Peer evaluation in practice. *ELT Documents, 131*, 119–25.

MANLEY, J. H. (1995). Assessing students' oral language: one school district's response. *Foreign Language Annals, 28*(1), 93–102.

MCKAY, P. (2000). On ESL standards for school-age learners. *Language Testing, 17*(2), 185–214.

MCKAY, P., HUDSON, C. & SAPUPPO, M. (1994). ESL bandscales, *NLLIA ESL development: language and literacy in schools project.* Canberra: National Languages and Literacy Institute of Australia.

MCKAY, P. & SCARINO, A. (1991). *The ESL Framework of Stages.* Melbourne: Curriculum Corporation.

MCNAMARA, T. (1998). Policy and social considerations in language assessment. *Annual Review of Applied Linguistics, 18*, 304–19.

MCNAMARA, T. F. (1995). Modelling performance: opening Pandora's box. *Applied Linguistics, 16*(2), 159–75.

MCNAMARA, T. F. & LUMLEY, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*(2), 140–56.

MESSICK, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

MESSICK, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–56.

MILANOVIC, M. (1995). Comparing language qualifications in different languages: a framework and code of practice. *System, 23*(4), 467–79.

MOELLER, A. J. & RESCHKE, C. (1993). A second look at grading and classroom performance: report of a research study. *Modern Language Journal, 77*(2), 163–9.

MOORE, T. & MORTON, J. (1999). Authenticity in the IELTS academic module writing test: a comparative study of task 2 items and university assignments. In R. Tulloh (Ed.), *IELTS Research Reports 1999* (Vol. 2, 64–106). Canberra: IELTS Australia Pty Limited.

MUNDZECK, F. (1993). Die Problematik objektiver Leistungsmessung in einem kommunikativen Fremdsprachenunterricht: am Beispiel des Französischen. *Fremdsprachenunterricht, 46*(8), 449–54.

NLLIA (NATIONAL LANGUAGES AND LITERACY INSTITUTE OF AUSTRALIA) (1993). *NLLIA ESL Development: Language and Literacy in Schools*, Canberra: National Languages and Literacy Institute of Australia.

NEIL, D. (1989). Foreign languages in the National Curriculum – what to teach and how to test? A proposal for the Languages Task Group. *Modern Languages, 70*(1), 5–9.

NORTH, B. & SCHNEIDER, G. (1998) Scaling descriptors for language proficiency scales. *Language Testing, 15* (2), 217–62.

NORTON, B. & STARFIELD, S. (1997). Covert language assessment in academic writing. *Language Testing, 14*(3), 278–94.

OSCARSON, M. (1984). *Self-assessment of foreign language skills: a survey of research and development work.* Strasbourg, France: Council of Europe, Council for Cultural Co-operation.

OSCARSON, M. (1989). Self-assessment of language proficiency: rationale and applications. *Language Testing, 6*(1), 1–13.

OSCARSON, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *Language testing and assessment* (Vol. 7, 175–87). Dordrecht, The Netherlands: Kluwer Academic Publishers.

PADILLA, A. M., ANINAO, J. C. & SUNG, H. (1996). Development and implementation of student portfolios in foreign language programs. *Foreign Language Annals, 29*(3), 429–38.

PAGE, B. (1993). The target language and examinations. *Language Learning Journal, 8*, 6–7.

PAPAJOHN, D. (1999). The effect of topic variation in performance testing: the case of the chemistry TEACH test for international teaching assistants. *Language Testing, 16*(1), 52–81.

PEARSON, I. (1988). Tests as levers for change. In D. Chamberlain & R. Baumgardner (Eds.), *ESP in the classroom: Practice and evaluation* (Vol. 128, 98–107). London: Modern English Publications.

PEIRCE, B. N. & STEWART, G. (1997). The development of the Canadian Language Benchmarks Assessment. *TESL Canada Journal, 14*(2), 17–31.

PLAKANS, B. & ABRAHAM, R. G. (1990). The testing and evaluation of international teaching assistants. In D. Douglas (Ed.), *English language testing in U.S. colleges and universities* (68–81). Washington D.C.: NAFSA.

PUGSLEY, J. (1988). Autonomy and individualisation in language learning: institutional implications. *ELT Documents, 131*, 54–61.

REA-DICKINS, P. (1987). Testing doctors' written communicative competence: an experimental technique in English for specialist purposes. *Quantitative Linguistics, 34*, 185–218.

REA-DICKINS, P. (1997). So why do we need relationships with stakeholders in language testing? A view from the UK. *Language Testing, 14*(3), 304–14.

REA-DICKINS, P. & GARDNER, S. (2000). Snares or silver bullets: disentangling the construct of formative assessment. *Language Testing, 17*(2), 215–43.

READ, J. (1990) Providing relevant content in an EAP writing test, *English for Specific Purposes* 1, 243–68.

REED, D. J. & HALLECK, G. B. (1997). Probing above the ceiling in oral interviews: what's up there? In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment.* Jyväskylä: University of Jyväskylä.

RICHARDS, B. & CHAMBERS, F. (1996). Reliability and validity in the GCSE oral examination. *Language Learning Journal, 14*, 28–34.

ROSS, S. (1998). Self-assessment in second language testing: a meta-analysis of experiential factors. *Language Testing, 15*(1), 1–20.

ROSSITER, M. & PAWLIKOWSSKA-SMITH, G. (1999). The use of CLBA scores in LINC program placement practices in Western Canada. *TESL Canada Journal, 16*(2), 39–52.

ROY, M.-J. (1988). Writing in the GCSE – modern languages. *British Journal of Language Teaching, 26*(2), 99–102.

SCIARONE, A. G. (1995). A fully automatic homework checking system. *IRAL, 33*(1), 35–46.

SCOTT, M. L., STANSFIELD, C. W. & KENYON, D. M. (1996). Examining validity in a performance test: the listening summary translation exam (LSTE). *Language Testing, 13*, 83–109.

SHAMEEM, N. (1998). Validating self-reported language proficiency by testing performance in an immigrant community: the Wellington Indo-Fijans. *Language Testing, 15*(1), 86–108.

SHOHAMY, E. (1993). *The power of tests: The impact of language tests on teaching and learning* NFLC Occasional Papers. Washington, D.C.: The National Foreign Language Center.

SHOHAMY, E. (1997a). Testing methods, testing consequences: are they ethical? *Language Testing, 14*(3), 340–9.

SHOHAMY, E. (1997b). Critical language testing and beyond, plenary paper presented at the American Association for Applied Linguistics, Orlando, Florida. 8–11 March.

SHOHAMY, E. (2001a). *The power of tests*. London: Longman.

SHOHAMY, E. (2001b). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–92.

SHOHAMY, E., DONITSA-SCHMIDT, S. & FERMAN, I. (1996). Test impact revisited: washback effect over time. *Language Testing, 13*(3), 298–317.

SHORT, D. (1993). Assessing integrated language and content instruction. *TESOL Quarterly, 27*(4), 627–56.

SKEHAN, P. (1988). State of the art: language testing, part I. *Language Teaching*, 211–21.

SKEHAN, P. (1989). State of the art: language testing, part II. *Language Teaching*, 1–13.

SPOLSKY, B. (1997). The ethics of gatekeeping tests: what have we learned in a hundred years? *Language Testing, 14*(3), 242–7.

STANSFIELD, C. W. (1981). The assessment of language proficiency in bilingual children: An analysis of theories and instrumentation. In R. V. Padilla (Ed.), *Bilingual education and technology*.

STANSFIELD, C. W., SCOTT, M. L. & KENYON, D. M. (1990). *Listening summary translation exam (LSTE) – Spanish* (Final Project Report. ERIC Document Reproduction Service, ED 323 786). Washington DC: Centre for Applied Linguistics.

STANSFIELD, C. W., WU, W. M. & LIU, C. C. (1997). *Listening Summary Translation Exam (LSTE) in Taiwanese, akak Minnan* (Final Project Report. ERIC Document Reproduction Service, ED 413 788). N. Bethesda, MD: Second Language Testing, Inc.

STANSFIELD, C. W., WU, W. M. & VAN DER HEIDE, M. (2000). A job-relevant listening summary translation exam in Minnan. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (Studies in Language Testing Series, Vol. 9, 177–200). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

TARONE, E. (2001). Assessing language skills for specific purposes: describing and analysing the 'behaviour domain'. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. F. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: essays in honour of Alan Davies* (Studies in Language Testing Series, Vol. 11, 53–60). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

TAYLOR, C., KIRSCH, I., EIGNOR, D. & JAMIESON, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*(2), 219–74.

TEASDALE, A. & LEUNG, C. (2000). Teacher assessment and psychometric theory: a case of paradigm crossing? *Language Testing, 17*(2), 163–84.

TESOL (1998). *Managing the assessment process: A framework for measuring student attainment of the ESL standards*. Alexandria, VA: TESOL.

TRUEBA, H. T. (1989). *Raising silent voices: educating the linguistic minorities for the twenty-first century*. New York: Newbury House.

VAN EK, J. A. (1997). The Threshold Level for modern language learning in schools. London: Longman.

VAN ELMPT, M. & LOONEN, P. (1998). Open questions: answers in the foreign language? *Toegepaste Taalwetenschap in Artikelen, 58*, 149–54.

VANDERGRIFT, L. & BELANGER, C. (1998). The National Core French Assessment Project: design and field test of formative evaluation instruments at the intermediate level. *The Canadian Modern Language Review, 54*(4), 553–78.

WALL, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13*(3), 334–54.

WALL, D. (2000). The impact of high-stakes testing on teaching and learning: can this be predicted or controlled? *System, 28*, 499–509.

WALL, D. & ALDERSON, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*(1), 41–69.

WATANABE, Y. (1996). Does Grammar-Translation come from the Entrance Examination? Preliminary findings from classroom-based research. *Language Testing, 13*(3), 319–33.

WATANABE, Y. (2001). Does the university entrance examination motivate learners? A case study of learner interviews. Akita Association of English Studies (ed.). Trans-equator exchanges: A collection of acadmic papers in honour of Professor David Ingram, 100–10.

WEIR, C. J. & ROBERTS, J. (1994). *Evaluation in ELT*. Oxford: Blackwell Publishers.

WELLING-SLOOTMAEKERS, M. (1999). Language examinations in Dutch secondary schools from 2000 onwards. *Levende Talen, 542*, 488–90.

WILSON, J. (2001). *Assessing young learners: what makes a good test?* Paper presented at the Association of Language Testers in Europe (ALTE) Conference, Barcelona, 5–7 July 2001.

WINDSOR, J. (1999). Effect of semantic inconsistency on sentence grammaticality judgements for children with and without language-learning disabilities. *Language Testing, 16*(3), 293–313.

WU, W. M. & STANSFIELD, C. W. (2001). Towards authenticity of task in test development. *Language Testing, 18*(2), 187–206.

YOUNG, R., SHERMIS, M. D., BRUTTEN, S. R. & PERKINS, K. (1996). From conventional to computer-adaptive testing of ESL reading comprehension. *System, 24*(1), 23–40.

YULE, G. (1990). Predicting success for international teaching assistants in a US university. *TESOL Quarterly, 24*(2), 227–43.

ZANGL, R. (2000). Monitoring language skills in Austrian primary (elementary) schools: a case study. *Language Testing, 17*(2), 250–60.