ED 365 145 FL 021 758

AUTHOR Alderson, J. Charles

TITLE Language Testing in the 1990s: How Far Have We Come? How Much Further Have We to Go?

PUB DATE 91

NOTE 27p.; In: Sarinee, Anivan, Ed. Current Developments in Language Testing. Anthology Series 25. Paper presented at the Regional Language Centre Seminar on Language Testing and Language Programme Evaluation (Ap:il 9-12, 1990); see FL 021 757.

PUB TYPE Reports - Evaluative/Feasibility (142) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Computer Assisted Testing; Conferences; Educational Trends; Financial Support; Futures (of Society); Intellectual Disciplines; Language Proficiency; *Language Tests; Research Needs; Second Language Learning; *Second Languages; *Testing; *Testing Problems; *Test Reliability; *Test Validity; Trend Analysis

ABSTRACT

A discussion of trends and progress in language testing looks at movement in the field since 1980, based on the themes and content of national and international conferences; trends in test content, method, and analysis; and work on the nature of proficiency and of language 'earning. It is proposed that movement evidenced by conferences is .argely sideways and backward; that while improvements have been made in test content, method, and analysis, there is little evidence that these improvements represent real advancements; and that research on the nature of proficiency and of language learning is still in its early stages. Four main reasons are given for the lack of progress: the relative youth of the discipline; dearth of replication, teamwork, and agenda in research; inadequacy of funding; and lack of a coherent framework or model. Areas in which attention will be important in the next decade are outlined, including: research on language learning; the washback effect of testing; validity of test content; knowledge of the structure of language proficiency; computer- based language testing and the impact of technology on testing; learner-centered testing; the role of judgment in language testing; and traditional concerns about test validity and reliability. (MSE)

LANGUAGE TESTING IN THE 1990S: HOW FAR HAVE WE COME? HOW MUCH FURTHER HAVE WE TO GO?

J Charles Alderson

INTRODUCTION

The metaphor I have chosen for my title relates to the notion of distance, of movement from A to B, or to C or Z. Reviews of language testing often employ a growth metaphor: papers and volumes are often called things like Developments in Language Testing. The implication is that language testing grows, like an organism, and in developing, changes from immaturity to maturity. In many ways, that is a useful analogy: certainly it is possible to look back over the recent history of language testing, and to claim that testing "has come of age". The specialisation has become a discipline. From being very much the junior partner, not to say ugly sister in language teaching and applied linguistics, language testing has "developed" to the point where it is almost respectable - almost an adult. Testers show, I believe, increased self-confidence within applied linguistics, display less of an inferiority complex, and show a much greater willingness to contribute from testing to related areas - teaching, second language acquisition, evaluation, and so on, as we shall see. I used to apologise for being a language tester, and tell strangers at cocktail parties that I was an applied linguist, or a teacher trainer. No longer. I even find that my non-tester colleagues are becoming interested in language testing, which they feel has become a lively, interesting area of research and development activity. Some have even expressed an interest in learning about how one goes about measuring the outcomes from language learning, how language proficiency might be identified, and how it is structured. So testing has developed and matured to the point where it is no longer dependent on mother disciplines and areas of interest. Testing has become acknowledged as a legitimate area of intellectual enquiry and professional engagement. At the 12th annual ILTRC in 1990 in San Francisco, there was talk of forming a professional organisation of language testing specialists. This is a sure sign that the field has matured, gained in confidence and is able to walk without holding its mother's hand, and is a development very much to be welcomed.

However, there is a drawback in the analogy with human development. Humans not only mature, they get older, and eventually wither and die. So if we

[redacted]

pursued the analogy, we would have to think about how and when language testing might enter middle and then old age, and what it could look forward to in the future. If testing is mature now, is it soon going to enter into decline and decay? Even if language testing is still in its adolescence, and has maturity and middle age to look for-ard to, the metaphor still implies the inevitability of decay and death. It is perhaps interesting to note that I once compared the life of a test to that of a human, suggesting that its existence was typically of the order of fifteen years:

> "After considering developments in English as a Foreign Language (EFL) .. I conclude that a test has a fairly well established life cycle of twelve to fifteen years. Once born, a test needs time and careful nurturing to develop, to attract more and more attention and testees, to establish credibility as an instrument for a particular purpose, to become recognized as valid and reliable .... This period seems to take between three and five years.
>
> Once established the test is then regarded as acceptable for a reasonable period of time. During this period it might be accepted by a variety of institutions, referred to in the testing and later teaching literature. It might be taken by large numbers of students, often forming the goal of their instruction and aspirations. This period might last anywhere between five and eight years. Towards the end of this period, however, signs of senescence appear in the shape of increasing criticism of the test's influence on teaching and on students' ambitions and lives .... Pressure may then build up within the test producing body itself ... for a change in test specification, test content, test format.. It may be that the test no longer fulfils its original function. Change may be instituted by academic applied linguists... or by the examination body itself, ... or it may be brought about by direct rather than invited teacher involvement. Whoever the agent of change, however, rebirth is then inevitable, usually after a gestation period of two to three years. And so we have another innovation: another baby test. However, the baby may have a very close resemblance to the parent, or it may look very different indeed from its predecessor" (Alderson, 1986, pp96-97)

However, although this may be true of a test, I do not believe it to be true of testing. Certainly there are people within language teaching who wish that it were true: that testing would decline and die, and that teachers and learners could then go about their daily lives unencumbered by tests. There is, after all, considerable resentment against tests, and their influence, and teachers in particular bemoan the washback effect. Many teachers also believe that they know what their learners have learned and have not learned, and how proficient or otherwise they are. The implication is that because they know their learners, they do not need tests, nor do they believe the information tests provide when it is different from what they themselves believe. Clearly language learning is a complex matter, as is the nature of the proficiency towards which learners are striving. Equally clearly a language test is going to be only a small, probably inadequate sample of what any learner has achieved or can do with language. And so many teachers, and some learners, criticise tests for being "unrepresentative" or even "misleading". Such teachers would be happy to see testing die. So would those who feel that testing not only constrains the syllabus, but also unduly restricts opportunities for learners, especially those learners who perform less well on the tests - who "fail" them.

However much such people hope that testing will die, their hopes are unlikely to be realised. Tests will probably always be needed as long as society is obliged to make selection choices among learners (as in the case of universitv entrance, for example), or as long as there is doubt about the validity or accuracy of other estimates of what learners might have learned (such as teacher judgements). Similarly, learners themselves frequently express the need for insight into their learning and achievement through tests. They want to know how well they are doing with reference to other people (norm-referencing) and how well they are doing with respect to the language, or some more or less absolute standard of what they need to achieve. And, of course, it is the professional responsibility of teachers and testers to provide that information, to the best of our ability.

So rather than hoping that tests will just go away, it is more realistic to try to improve the tests that we have, so that negative washback can become positive, so that tests reflect what learners and teachers think learners have learned and can do, and so that the decisions that are made on the basis of test results are as fair and reasonable as they can possibly be. Which is why reliability and validity are so important, and why it is important that publicly available language tests and examinations should meet clearly established and accepted standards. One of the first tasks that the proposed association of language testing specialists will need to address, is that of standards: what represents good practice in language testing, how is it be identified, fostered and maintained? What are the existing standards of our examining bodies and language tests, and should and can these standards be improved? So one task for the 1990s will certainly be to improve on the growing professionalism of language testers and language tests, and to set standards to which tests might - should - must - aspire.

However, this notion of aspiration suggests a different metaphor from that of growth, development, maturation, then decline, decay and death. It suggests aspiring to a goal, something distant that is getting closer, or high that is getting nearer. Hence the metaphor of distance contained in my title, and the idea it suggests of progress over distance. I am interested in running, and have done quite a few long-distance events - half marathons, marathons, and ultra events. I find short-distance running - sprinting - very painful, and do not enjoy it. Which partly explains why I do long-distance running, and it may also explain why I enjoy testing, or at least do not find it frustrating. You need stamina to be a long-distance runner, you need stamina to develop good tests, you need stamina to do language testing research, you need stamina and patience to see improvements in language testing over time. The language testing run - the distance we have to cover - is long, very long. Language learning is very complex, and there is a great deal we do not know about it: second language acquisition researchers are really only at the beginning of their journey of discovery. Language proficiency is a complex phenomenon, and is very little understood, despite the best efforts of many social science disciplines to attempt to elucidate it. There are many different, and indeed sometimes competing models of language proficiency, and we are barely at the beginning of operationalising and testing and validating those models. Research into testing methods is fairly recent, and has a long way to go yet: there is a lot we do not know, a frightening amount of ground to be covered. We may have made some progress, as I shall discuss, but there certainly is a lot more to be made.

HOW FAR HAVE WE COME? HOW MUCH FURTHER HAVE WE TO GO?

My title is intended to be suggestive of a range of topics in and related to language testing.

Firstly it asks about progress in language testing - has there been any? Are language tests any better now than they used to be? Have we now achieved a greater understanding of what the problems in language testing are, or how they might more appropriately be approached if not solved? Have we moved forwards at all, or have we been "running on the spot" these last ten years and more? Do we now understand more about the nature of language proficiency: its structure, how it is acquired or lost? Do we have a better idea of what aptitude for language learning is? Do we now know more about how learners learn, and how we can best measure that learning?

And secondly, the question can be seen as coming, not from a language tester, but from a learner: *How far have I come: what have I learned so far? What progress have I made? What is my achievement?* as well as *How much further have I to go: what is my ultimate goal? What is the nature of language proficiency?*

I shall be arguing that we need to bear in mind that we need to be concerned not only with testing, but with tests; not only with the nature of second language proficiency, and the design and researching of language proficiency tests, but also with language learning, and the design and researching of achievement tests; not only with testers, and the problems of our professionalism but also with testees, with students, and their interests, perspectives and insights. As I said at a conference in Sri Lanka five years ago, "testing is too important to be left to testers".

ACTIVITY AND PROGRESS

What progress has there been, then, in language testing? Certainly, if we stick with the running metaphor, there has been a great deal of movement and activity.

Since 1980, language testing has indeed been moving apace. We now have an internationally respected journal **Language Testing**; newsletters like **Language Testing Update** and **Language Testing Notes**; a Special Interest Group within IATEFL for Language Testing; an annual Language Testing Research Colloquium; and many publications in language testing. Where in 1980 we had relatively few publications specifically in language testing, now we have many. Peter Skehan's recent survey article lists 215 publications in the Bibliography, of which only thirty-five were published before 1980. Grant Henning's book **A Guide to Language Testing** has been recently complemented by Lyle Bachman's volume on **Fundamental Considerations iu Language Testing**. The second edition of Brian Heaton's volume: **Writing English Language Tests** has been complemented by Arthur Hughes' book **Testing For Language Teachers**, and Cyril Weir's **Communicative Language Testing**. The Language Testing Research Colloquium itself has given rise to several volumes including Jones et al (1985) from the 1983 colloquium and Stansfield (1986) from the 1985 colloquium, on Technology and Language Testing. In Europe, the Interuniversitare Sprachtestgruppe (IUS) organised annual conferences which gave rise to several volumes on the theme of **"Practice and Problems in Language Testing"**.

The list is much longer than this brief, unrepresentative selection. It should, however, be sufficient to illustrate the fact that language testing appears to have generated a lot of activity, of work, of meetings, publications, discussions, seminars, courses, even tests.

Yet where have we got to? How far have we advanced? What do we now know that we did not know in 1980? It is instructive, I believe, to compare what w.is happening in 1980 with what appears to be happening in 1990.

To take an international example, first. The International Language Te.ting Research Colloquium, which held its first meeting in 1978, this year held its 12th annual meeting. The theme of the 1980 meeting in San Francisco was: The Construct Validation of Oral Proficiency Tests. Ten years later, in San Francisco again, the theme was: A New Decade in Language Testing: Collaboration and Cooperation. In 1980, the colloquium concentrated on oral testing, with papers on oral proficiency scales, the interviewer's role, convergent discriminant validation of oral and written tests, and above all extended discussion of Bachman and Palmer's pilot work on the construct validation of tests of speaking and reading. The reader may recall that the upshot of the Bachman-Palmer study was that speaking was shown to be measurably different from reading, but that there was evidence of method effects.

In 1990, the colloquium was much larger - 106 people attending compared with 29 invited participants in 1980. Partly as a consequence, there was a much greater variety of papers and topics covered. It is difficult to summarise the topics without distortion, but my subjective impression is that the papers fell roughly into four areas:

i) test methods: discussion of various ways to measure traits, including the effect of the prompt in writing tests, comparisons of open-ended and multiple choice techniques, and the effect of instructions in summarising tests

ii) test analysis: the use of Item Response Theory

iii) test content: field specificity of speaking tests, specificity of tests of reading skills, ESP test content and test bias, approaches to the validation of reading tests

iv) test development and test analysis: a colloquium on the TOEFL - Cambridge Comparability study, and another on the development of the new IELTS test, from the point of view of the role of grammar, the nature of the listening and speaking tests, and the issue of subject-specific testing

Yet although clearly more varied, it is not clear to me that the 1990 colloquium was an advance on the 1980 one. In some ways, it was a step backwards, since in 1980 there was a common theme, with research papers bearing on the same issue from a variety of angles, and potentially throwing light on problems that might have been said to persist in oral testing. However, many of the problems that were aired in 1980 are still current: the issue of oral proficiency scales is eternal, was, for example, addressed in the recent ELTS Revision Project, and we will hear something about this at this conference from David Ingram.

To turn to a national example. Almost exactly ten years ago, in May 1980, Alan Davies and I hatched a plan to hold an invitational conference at Lancaster with the aim of reviewing developments and issues in language testing. After some deliberation and discussion, we agreed that the three main "issues" of interest to British language testers, and hopefully more widely also, were: communicative language testing; testing English for specific purposes; the unitary competence hypothesis: testing general language proficiency. The results of the conference were eventually published as "Issues in Language Testing" (Alderson and Hughes 1981); out of that meeting also came the **Language Testing Newsletter**, which eventually became the journal **Language Testing** and at the same conference, discussions were held which led to the Edinburgh ELTS Validation Study. I think we had a very definite sense that we were at the beginning of interesting developments, and that much could happen. A subsequent follow-up conference was held at Reading University on the same three topics, and the proceedings were published as **Current Developments in Language Testing** (Hughes and Porter, 1983).

At the end of 1989, the Special Interest Group in Testing within IATEFL organised a Testing Symposium in Bournemouth, entitled Language Testing in the 1990s: The Communicative Legacy. The symposium attracted a variety of presentations from examination bodies in the United Kingdom, from teachers involved in testing, and from testing researchers. In addition, a mini-colloquium took place, where precirculated papers were reacted to by invited speakers. The proceedings are about to be published in the ELT Documents series: the main themes centre around three areas: oral testing, computer based testing, testing and teaching. In the mini-colloquium, the papers concentrated on communicative testing and the role of grammar; resistance to change; and the role of judgements, and band scales, in proficiency assessment. Many of these themes are by now probably familiar from the previous meetings I have mentioned: communicative testing and oral testing in particular, but also the relationship between teaching and testing, and the nature of proficiency. The newcomer is computer-based testing, and I shall come back to that topic shortly. In general, however, I did not and do not get the impression that we have been building upon previous research and previous discoveries and successes over the past decade, and in this impression I am strengthened by Peter Skehan, who is not only the author of an excellent survey article of language testing (Skehan, 1988), but also presented the opening overview paper at the Bournemouth symposium. In his paper, Skehan claims that there has been little notable progress in testing in the past decade, which he attributes in part to conservative forces within society as well as within testing. Firstly, significant advances in testing tend to depend upon research using large batteries of tests which require large numbers of students. These are not readily available, and certainly require considerable investment of resources: time, money, people - in order to exploit the possibilities. Secondly, testing has long recognised the need for instruments to be reliable, and since, at least by traditional statistical measures, it has been demonstrated that established tests and techniques like the multiple choice technique can be highly reliable, there must be an inbuilt resistance to change for tests and techniques that may be less reliable. Thirdly, the time needed to develop new tests for public administration is at least two, usually more like three years, and since such innovation will usually entail changes in syllabuses, teaching materials and teacher training, there is an inbuilt resistance to changing systems through innovation. (An interesting exception and example of this is provided by the recent ELTS Revision Project of which I was Project Director, where although we were charged with innovating in test content and test method, we were also told very firmly that the new test must be consistent with the old test, and should not involve radical departures from existing practice!) Fourthly, examining bodies tend to be obsessed with security, and therefore are understandably very reluctant to allow researchers access to their probably very considerable datasets. If it were possible to explore such datasets, and to compare them, we might well be in a much better position to understand what our current measures do and do not achieve, and to recommend research and development programmes which could contribute to progress in our understanding of what and how to test. Fifthly, Skehan points out that language testing has not, on the whole, been well served by linguistic or applied linguistic theory. Linguistics, especially that branch of linguistics that is concerned with Universal Grammar, Skehan dismisses as being irrelevant to testing. He argues, and I tend to agree, that linguistics has provided little or nothing in the way of insights and understandings that can lead or has led to improvements or even changes in language test content. Sociolinguistics, and to some extent psycholinguistics, at least as represented by second language acquisition studies, have indeed made some progress, and interesting ideas are beginning to suggest themselves to language testers. Yet even in these fields, the profusion of competing and contradictory models, often with very slim empirical foundations, inhibits the language tester or applied linguist from selecting " the best model" on which to base his or her language test. Again, the ELTS Revision Project is a good case point. The previous ELTS test had been based upon John Munby's model for syllabus design - the communicative needs processor (Munby, 1978). Brendan Carroll and his associates took the model and appear to have applied it to the design of test specifications and test content. The Revision Project was asked to re-examine this, on the grounds that the Munby model was oldfashioned, out of date, and needed to be replaced. So one of our tasks was to identify a model of language proficiency on which our test should or could safely be based. Alderson and Clapham (1989) report on the results of our attempts to identify an applied linguistic model to replace Munby: we failed. There was general consensus that Munby was indeed no longer appropriate, but absolutely no agreement on what might replace Munby. In the end, we were forced to design our own construct, being as eclectic and openminded as we possibly could in the circumstances.

The point of this anecdote is simply to reinforce the important point that Skehan makes: language testing has not been well served by applied linguistic theory, and has been forced to reach its own solutions and compromises, or to make its own mistakes. Indeed, it is significant, in my view, that the most likely influential model of second language proficiency to emerge at the end of the 1980s is the model proposed by Lyle Bachman in his 1990 book. I shall return to this model shortly, but what is significant here is the fact that Bachman is a language tester, long concerned with researching the nature of language proficiency. It is on the basis of his research and empirical findings, coupled with his experience in language testing, and his consideration of issues like test method effects and the facets, as he calls them, of test design, that his model has developed. Clearly his model owes a dept to Canale and Swain, and Canale's later modifications, which model itself clearly has its origins in much sociolinguistic thought, but as Skehan points out, it is surely significant that the model had to be elaborated by a language tester, and is now beginning to be operationalised through work on the TOEFL-Cambridge comparability Study - of which more later.

So, in the United Kingdom as well as internationally, the impression I gain is that although there has been much movement, a lot of it is sideways and backwards, and not much of it is forwards. Are we going round and round in circles?

What about progress closer to hand? What does a comparison of the RELC Seminars on testing and evaluation in 1980 and 1990 reveal? In 1981, John Read edited the proceedings of the 1980 RELC conference on testing and evaluation: **Directions in Language Testing**. In this volume, the main themes were: The Validation of Language Tests; The Assessment of Oral Proficiency; The Measurement of Communicative Competence; The Cloze Procedure: New Perspectives; Self-Assessment of Language Proficiency; The Interaction of Teaching and Testing. With the possible exception of the topic of the cloze procedure, who would wish to deny that such themes might not be just as appropriate to the 1990 RELC Seminar? Communicative language testing appears on the programme of this seminar in many disguises. The assessment of oral proficiency is still topical, as is self assessment, the relationship between teaching and testing, test development and test validation. I shall be very interested as I listen to the many presentations at this seminar to see whether we are building upon previous research and understandings, whether we are reinventing wheels that have already been shown to be adequate, or even worse attempting to produce new wheels that are more square than the old ones. Will we, like our reinvented wheels, continue to go round and round in circles, possibly bumping as we go, thanks to the irregularities in the circles? That, I believe with Peter Skehan, is the challenge to language testing in the 1990s: can we make progress, can we go forward, or must we continue to go round in circles? One of the major contributions that this RELC Seminar could make to language testing, I believe, is to help us to answer the questions in my title, which I believe to be central. What progress have we made made in language testing : What do we now know, that we did not know in 1980? How far have we come? And secondly, what do we still not know? What needs to be done?

If by the end of the Seminar we have got our personal answers to these questions, and if we have also a sense of an emerging consensus among language testers of the answers, then we will not only have achieved a great deal, but will also be in a position to move forward.

I do not wish to pre-empt your own thoughts on progress to date and the need for further progress, but I fear that it would be remiss of me, given my title and the expectations it may have aroused, not to offer my thoughts on what we have and have not achieved to date. However, I shall do so only briefly, as I do hope that this seminar will clarify my own thinking in this area. Nevertheless, let me at least indicate the following areas of progress and lack of progress:

TEST CONTENT

The last ten years have seen apparent improvement in the content of language tests, especially those that claim to be "communicative". Texts are more "authentic", test tasks relate more to what people do with language in "real - life", our tests are more performance related and our criteria for evaluating performance are more relevant to language use. Advances in performance testing have accompanied a movement away from "discrete-point", knowledge-focussed tests and have benefited from more integrative approaches to language assessment. However, we do not know that this apparent improvement is a real advance in our ability to measure language ability. What analyses and comparisons have been done suggest that there is little detectable difference between the "new" and the "old", and much more work needs to be done to establish that progress has indeed been made. Lyle Bachman and his co-workers have developed a useful instrument for the TOEFL-Cambridge Comparability Study, intended to identify and examine aspects of "Communicative Language Ability" (based upon the Bachman model, Bachman 1990), but even this instrument when perfected will only enable us to compare tests in content terms. Much more work will need to be done before we can relate the empirical analysis of test performance to an examination of test content. We are still very far from being able to say "I know how to test grammatical competence", "I know how to test the ability to read in a foreign language", and so on.

TEST METHOD

Research has clearly shown that there is such a thing as method effect in testing. Given that we are not interested in measuring a person's ability to take a particular type of test, it is important that we minimise method bias in our measures and maximise their ability to measure trait. We are more aware than we were that there is no such thing as the one best method, much less one best test. We have, I believe, generally accepted that no one test method can be thought of as superior, as a panacea. There have been interesting attempts to devise new testing methods for particular purposes, which remain to be validated, but which offer alternatives to the ubiquitous multiple-choice. We have as yet no means, however, of estimating the method effect in a test score, much less any way of predicting test method effects or of relating test method to test purpose. The development of an instrument to identify possible test method facets through the TOEFL-Cambridge Comparability Study is a welcome contribution to our future understanding of this issue.

TEST ANALYSIS

Considerable advance has been ma'e in the quantitative and qualitative tools available to testers. The increased use of Item Response Theory, of multivariate statistics and casual modelling, including confirmatory factory analysis, has contributed to an increased understanding of the properties of language tests, and recent developments in the use of introspective techniques for the analysis of test performance promise to increase further that understanding. At the same time, the use of more sophisticated techniques reveals how complex responses to test items can be and therefore how complex a test score can be. We have much more work to do before we can claim to have identified what it is that any test tests, and therefore are equally far from being able to claim that we have perfected our tools for test analysis.

THE NATURE OF PROFICIENCY

We now know, or believe, that the answer to the question: what is language proficiency? depends upon why one is asking the question, how one seeks to answer it, and what level of proficiency one might be concerned with. It is generally accepted that the UCH was overstated, and that proficiency consists of both general and specific components. We know that speaking can be different from reading. We also now know, thanks to the work of Gary Buck (Buck, 1990), that reading and listening can be empirically separated, provided that certain conditions are met. We also know from various sources, including the empirical work of Vollmer, Sang et al, and Mike Milanovic in Hong Kong, that the nature of proficiency depends upon the level of proficiency. Advanced learners tend to exhibit a relatively integrated proficiency, which therefore tends to be unifactorial. Similarly, virtual beginners are likely to exhibit a non-differentiated proficiency. Whereas intermediate level learners - the majority - tend to show differentiated abilities across macro-skills, and therefore their test performance tends to be multifactorial. However, we are still at the beginning of this sort of research, and many carefully designed studies with large test batteries, homogeneous subjects, and adequate background information on learning histories and biodata will be needed before we can be more definitive in our statements about the nature of language proficiency than that.

THE NATURE OF LANGUAGE LEARNING

We know that language learners do learn language. We know very little about how, about how long it takes, under what conditions, and therefore we know little about how to measure the progress of learners. We do know that proficiency tests are very insensitive to such progress (as Palmer and Des Brisay will tell us at this conference). We lack more sensitive measures, and we are therefore unable even to make useful statements about the impact on learning of one method/teacher /classroom/syllabus over any other. We ought as a profession to knew more about language learning than that it is complex and unpredictable! Our previous tendency in language testing to concentrate on the development and validation of proficiency tests must be reversed in the next decade by a concerted effort to devise tests that are sensitive to learning on particular programmes, that will help us to explore the nature of language learning, and to contribute to second language acquisition studies, to programme evaluation, to language teaching and to applied linguistics more generally.

REASONS FOR LACK OF PROGRESS

So much for my personal view on our achievements and lack of achievements to date. Before I offer a few thoughts myself on the directions I believe we must go in language testing in order to make further progress, I should like to expand on Skehan's ideas at the Bournemouth symposium as to why there has been so little progress in testing, and then speculate on how this might be overcome.

There are, I believe, four main reasons:

i) Language testing is a young discipline, and has only been taken seriously, and taken itself seriously, in the past twenty years or so. As Alan Davies once pointed out, the resurgence, or increased attention and respect accorded to language testings, is due to the influence of people like John Oller, and latterly Lyle Bachman, but this is relatively recent.

ii) Language testing suffers from a lack of replication: this is a problem common to the social sciences. There is considerable pressure on researchers, especially doctoral students, to be "original". As a result, we do not yet have a tradition of team research, or established research agendas, which researchers in various centres are working on.

iii) The problem of lack of replication, lack of research teams, and lack of a research agenda is in part due to a lack of funding. It is relatively unusual, at least in the United Kingdom, and in my experience also elsewhere, for language testing - especially achievement testing - to receive funding from the research councils or other sponsors of social science research. As a result, the research that gets done tends to be done by docioral students, striving to be original, or by individual researchers working on their own (this is a point also made by Skehan), or by researchers within organisations like the Defense Language Institute in the USA, or the Centre for Applied Linguistics, who are necessarily bound by the priorities of their organisations. Occasionally the examination bod es fund outside research - the TOEFL - Cambridge Comparability Study is a good example of this , as are some of the research studies reported in the ETS TOEFL Research Report Series. However, understandably enough, the examination and test development bodies are more interested in funding research that will contribute directly to the development of their own tests and examinations, rather than to progress in language testing more generally.

iv) Lack of a coherent framework within which to work, so that ideas can contribute to each other, allowing the systematic exploration of one aspect and its relationship to other aspects of the framework or model.

In order to make progress in language testing, we need to pay attention to these problems. The first problem will, of course, resolve itself as testing becomes a more established field of study. There are reasons to believe, as I suggested at the beginning, that this has already occurred, that language testing is now sufficiently developed and mature, or language testers are sufficiently well trained and experienced, for this no longer to be a problem. The second problem can be overcome by an awareness of its existence: those of us who direct and supervise research should consciously encourage replication and the accumulation of research findings. We should also consciously seek opportunities for collaboration among researchers and research teams. The fourth problem - the lack of a common framework - is partly resolved by the appearance of the Bachman model, which is beginning to be welcomed by testing researchers in a variety of situations as a very useful and usable starting point and framework. I am sure that we would all have our individual misgivings or criticisms of parts of the model, but that should not prevent us from endeavouring to operationalise aspects of it, in order to explore relationships among them. The third problem is perhaps the most problematic: funding for research and development. I have no easy solutions to that, but will be very interested to hear what other have to say about this, from their institutional and national perspectives.

DIRECTIONS

In what follows, I offer a few of my own thoughts, on what I believe to be important areas for language testing to pay attention to in the next decade and more.

1. Achievement and research in language learning

Now that testing has come of age, it is time for testers to make major contributions to other areas of applied linguistics. Three related areas come to mind immediately: programme evaluation, second language acquisition research and classroom learning. In each case, it will be important for language testers to pay much more attention to the development and researching of achievement tests. In each case, what is needed is a set of carefully constructed, highly specific tests which can be shown to be sensitive to learning. The concentration of language testing researchers on developing and researching proficiency tests is understandable: most funding bodies want proficiency tests, rather than tests that relate reliably, and validly and directly to specific achievement on particular syllabuses, materials and programmes, or to the acquisition of particular language items/features/skills. The fact is that much programme evaluation uses inappropriate tests: either proficiency tests which can hardly be expected to be sensitive to learning in a detailed rather than global sense; or poorly constructed, untrialled or unvalidated "achievement" tests (the use of the cloze test in Canadian immersion studies, the Bangalore evaluation, etc is a case in point). The net effect is that findings are of dubious validity or indeed certainty, and our professional knowledge of the effect let alone effectiveness of teaching and learning is minimal. Put crudely, we ought to know far more than we do about the nature of language learning and I believe that one of the reasons we do not is the neglect of the contribution that language testing could make to gathering insights in this area. There are, of course, immensely difficult issues to be faced in deciding when and what someone has learned something, but these are not insuperable, and I believe that the 1990s will see much more collaboration between language testers, second language acquisition researchers, programme evaluators and language teachers than we have seen hitherto.

2. Washback

It is a commonplace to declare that tests have an impact on teaching - washback is everywhere acknowledged and usually deplored. At the same time, it is not uncommon to point out that tests can have both negative and positive influences on the curriculum, a fact which has been used in some settings in order to bring about innovation in the curriculum through the test. Usually, the test is said to lag behind innovations and progress in materials, teacher training and classroom practice, until the dissonance between the two becomes so uncomfortable that the test has to change. In some settings, however, deliberate innovation in the content and method of the examinations has been used to reinforce or in some occasions even go in advance of changes in materials and methods, However, in both sets of circumstances - where the test is held to have negative washback on teaching and where the test is being used to bring about classroom change - there is remarkably little evidence of the impact of the test. What there is is largely anecdotal, and not the result of systematic empirical research. What is needed, and there are signs that this will become an increased focus for testing related research in the future, is research into the impact of tests on classrooms. Do teachers "simply" use previous exam papers as textbook material? If so, do they simply expect students to take the tests, and then to receive the answers? How long in advance of the exam does such teaching begin, and what do students think of it , and how do they benefit from it? Why do teachers do it - if they do? Are there other, equally or more effective strategies for preparing students for exams, or for helping students to perform to the best of their ability in tests? What do tests do to the process of learning? Is washback necessarily negative, and do teachers necessarily and inevitably resent the influence and power of the test, or do they welcome it as providing motivation for learners and guidance to the teacher? Studies of the washback effect are only just beginning - Israel and Nepal are two examples, and my own University is involved in a four year project in Sri Lanka to seek to determine attitudes to tests and the nature of their impact.

3. Test Content

The last few years have seen a resurgence in interest in the content validity of tests. Over the past ten years there has developed a tendency for test developers to devise taxonomies of the skills and content being tested by their tests. Such taxonomies are typically contained in the Specifications of the test, which are used for the guidance of item writers, and they have been heavily influenced by the writings of curriculum and syllabus developers. The classic example of these in the USA is Benjamin Bloom and his associates, in the Taxonomy of Educational Objectives, and in the United Kingdom in EFL/ESP in the work of John Munby and his Communicative Needs Processor. The existence of taxonomies in test specifications has led to an attempt to test individual skills and objectives in individual items, and to the concomitant claim that certain items do indeed test certain skills/objectives.

Unfortunately, however, recent research has begun to cast doubt on these claims, at least to the extent that it has proved somewhat difficult in some circumstances to get "expert" judges to agree on what is being tested by individual items. If judges do not agree with each other, or with the test constructor, on what items are testing, then it becomes somewhat difficult to substantiate claims as to the content validity of an item, and conceivably also a test.

The development and use of introspective techniques in other areas of applied linguistics - especially in the study of reading - has led to their application to an understanding of what test candidates do when they are taking tests. Insights gained to date have centered upon test-taking strategies - what might be called test-wiseness and its absence: how students approach the task of taking a test - and test-processing: what students report of their mental processes when they are reading and listening, writing responses and completing multiple-choice grammar tests. What this new area of test analysis is beginning to show is that students approach test items in a highly individual way and, moreover, that students get the correct answer for a variety of different reasons. Sometimes they get the answer right without knowing the right answer, sometimes they get the answer wrong whilst clearly displaying the ability being tested. Even more worrisomely, in some ways, is the fact that individual students have been seen to get the answer right, yet have displayed abilities that were not supposedly being tested, nor have they displayed evidence of the ability that the test constructor believed was being tested.

If individuals respond to single items individually, revealing different skills and abilities in so doing, and if "expert" judges disagree about what is being tested by individual test items, then it is unclear whether we are justified (a) in saying that a given item is testing a particular skill for any group of learners, and (b) in grouping together the responses of different learners to the same item for the purposes of item analysis (even facility values and discrimination indices). If there are doubts about the legitimacy of grouping individual responses (which are at least potentially different psycholinguistically) to one item, there must also be doubts about the wisdom and indeed interpretability of grouping responses to items to arrive at test scores for one individual, and even less to arrive at group test results. Given that traditional test statistics - reliability indices and validity coefficients and calculations - depend upon grouping data - perhaps it is small wonder that factor analyses of large datasets of performance on large numbers of items result more frequently than not in unifactorial structures, or in mult-factorial views of proficiency that are difficult to interpret. This is at least an argument for interpreting statistical data cautiously, especially if it runs counter to our intuitions and insight into language learning from other perspectives. It is not an argument for doing away with language tests altogether.

4. Structure of language proficiency

The appearance of **Issues in Language Testing Research**, the admission by John Oller that he had pushed the Unitary Competence Hypothesis too far, and above all the careful empirical work of Bachman and Palmer led many to declare that the notion of "general language proficiency" had gone too far: language proficiency is both unitary and divisible at the same time, it seems. Thus there is a common or general factor in proficiency as measured by test results and also evidence for separable components, sometimes relating to the macro-skills, sometimes to less easily definable traits. For a while in the early 1980s, the debate was quiet, all was resolved, we thought. However, recently further research evidence for a very strong general factor as provided by researchers like Fred Davidson and the work of the Comparability Study has led some to reconsider their position. The to-some-disappointing finding that the factor structure of test batteries is more unifactorial than theory would lead us to expect is being accounted for by the notion that the nature of language proficiency may vary depending upon the level of proficiency. Thus advanced learners might be thought to have integrated their abilities in different skill areas, and therefore to manifest a general proficiency, or at least a proficiency that is relatively undifferentiated in the sense that no one skill is radically distinguishable from another skill. Good language users tend to be good at grammar and reading, writing and speaking, listening and vocabulary. Thus one might expect a unifactorial structure for language proficiency at the higher levels. However, lower intermediate students might well find that their reading abilities far outstrip their speaking or listening abilities, and therefore one might expect that at the lower levels of proficiency, language proficiency is more differentiated, more multi- factorial. Recent research in this general area does seem to offer some evidence to support this view, and it is likely that further research into the nature of language proficiency will have to look at more homogeneous groups than has been the case in the past. Grouping candidates from widely varying cultural, ethnic, linguistic and educational backgrounds together in order to make inferences about test content and construct and therefore also about language proficiency, is a dubious exercise at best, and also possibly highly misleading. Of course, researchers have used such heterogeneous populations partly because of the tests being investigated - especially the TOEFL - and also because of the populations from whom data has been gathered - typically the Language Institutes associated with an American University, whose populations may be relatively homogeneous educationally but certainly not culturally or linguistically.

It is hoped that the realisation of the problems associated with this, and the need for closer attention to the need to test language achievement rather than proficiency, might well lead to an increase in the studies that are conducted on populations in individual countries, within particular educational settings. This might also enable us to focus more clearly on achievement and learning within institutions and systems.

5. CBELT: computer-based language testing and the impact of technology.

One major development since 1980 has been the advent of personal computers utilising powerful and advanced micro-processors. Such computers are increasingly being used not only for the calculation of test results and the issuing of certificates, but also for test delivery and scoring. Computerised adaptive testing is an important innovation, where the computer "tailors" the test that any candidate takes to that candidate's ability level as revealed by his/her performance on previous test items. Thus on the basis of his/her response to the first item, the computer calculates the candidate's ability level (using IRT in some form) and selects the next item from an item bank at that estimated level of ability. Through an iterative process of estimation and administration, the computer is able to achieve a reliable estimate of a candidate's ability with fewer items than is normally possible, thus increasing the efficiency with which tests can be administered and reducing the time necessary.

Computers are also increasingly being used for the routine administration of a range of different tests for different purposes. Unfortunately, the tests administered are usually either multiple-choice, or fixed ratio cloze tests scored by the exact word procedure. Whereas such test methods are coming under increasing scrutiny and criticism elsewhere in language testing, the advent of the computer has to date proved to be a conservative force in test development: test methods are being used that might otherwise be questioned, because it is thought difficult for the computer to deal with other test methods. The tremendous opportunities that computers might offer for innovation in test method have not yet been taken up, despite the possibilities outlined in Alderson, 1988.

The speed, memory, patience and accuracy of the computer would appear to offer a variety of possibilities for innovation in test method that should be actively explored in the 1990s. In addition, however, it is already clear that delivering tests on computer allows the possibility for a blurring of the distinction between a test and an exercise. The computer can assess a student's response as soon as it has been made: this then allows the possibility for immediate feedback to the student before he/she progresses to the next item. It also allows the possibility of giving the student a "second chance", possibly for reduced credit. The computer can also provide a variety of help facilities to students: on-line dictionaries can be made easily available, and with not much more effort, tailor-made dictionaries - directly relevant to the meanings of the words in the particular context - can also be available, as can mother tongue equivalents, and so on. In addition, the computer can deliver clues to learners who request them. These clues can be specific to particular items, and can consist of hints as to the underlying rules, as to meanings, as to possible inferences, and so on. Again, the test developer has the possibility of allowing access to such clues only for reduced credit. Moreover, the computer can also offer the possibility of detailed exploration of a particular area of weakness. If a student performs poorly on, say, two items in a particular area, the machine can branch the student out into a diagnostic loop that might explore in detail the student's understanding and weaknesses/strengths in such an area. If thought desirable, it would be easy to branch students out of the test altogether into some learning routine or set of explanations, and then branch them back in, either when they indicated that they wished the test to continue, or once they had performed at some pre-specified criterion level.

In short, a range of support facilities is imaginable through computers - and indeed software already exists that allows the provision of some of these ideas. The provision of such support raises serious questions about the distinction between tests and exercises, and the consequences of providing support for our understanding of candidates' proficiency and achievement. Since the computer can also keep track of a student's use of such facilities, it is possible to produce very detailed reports of progress through a test and of performance on it, thus allowing the possibility of detailed diagnostic information. The big question at present is: can teachers and testers use this information? Will it reveal things about a student's proficiency or achievement or learning or test taking strategies that will be helpful? We do not yet know, but we now have the hardware, and partly the software, to find out, and a duty to explore the possibilities and consequences.

6. Learner-centered testing

The very real possibility of the provision of support during tests, and the tailoring of tests to students' abilities and needs, raises the important issue of student-centered testing. For a long time in language teaching there has been talk of, and some exploration of the concept of, learner-centered teaching and learning. This now becomes an issue in testing, and as teachers we will need to decide whether we need and want to explore the possibilities. Interest in students' self-assessment has continued throughout the decade; the advent of introspective methods for investigating test content and test taking processes allows us to gather information on test processes from a student's perspective and thus to get a different, student-centered, perspective on test validity. It is possible to envisage further developments where students are invited to contribute more directly to the test development process, by getting them to indicate what they consider suitable measures of outcomes from instructional programmes might be. What do they think they have learned during a programme and how do they think they can demonstrate such learning? An increased focus on such questions could help learners as well as teachers become more aware of the outcomes of classroom learning, which would in turn inform those who need to develop classroom progress and achievement tests.

Clearly such suggestions are revolutionary in many settings, and I am not necessarily advocating that students design tests for themselves, their peers or their successors, at least not immediately. It is often necessary to begin such a development cautiously: one way already suggested and indeed being tried out in various contexts is to ask students what they are doing when they respond to test items. Another way is to ask students to comment and reflect on the discrepancy between their test results or responses, and their own view of their ability, or their peers' views or their teachers views. Such explorations may well help us to understand better what happens when a student meets a test item, and that might help us to improve items. It might also help students to understand their abilities better, and might even encourage students to contribute more substantially to test development. The ideas may appear Utopian, but I would argue that we would be irresponsible not to explore the possibilities.

7. Judgments in language testing

Language testers have long been aware that testing is a judgemental activity. The development of multiple-choice tests was an attempt to reduce unreliability of scoring judgements, by making it possible to mark tests by machine. Those concerned with the development of direct or semi-direct tests of speaking and writing abilities have traditionally sought to establish the reliability of subjective scoring judgements/careful training through scorers, through inter and intra-rater comparisons, and it is common practice in many parts of the world to report scorer reliability coefficients. However, there are many areas beyond scoring where judgements are important in language testing. These include: the design of the test, involving decisions about test method and test content, and judgements about what is being tested by items and subjects. Testers also have to make judgements about the appropriacy of items to given target populations, especially important in settings where pre-testing and pilotting of tests is impossible, or massively difficult. Testers also often have to decide which students should be deemed successful on tests and which not: who has passed and who failed? Some traditions of language testing - I am thinking here especially of the British tradition - rely very heavily indeed on "expert judgements". Examination bodies select individuals to produce, edit and mark their tests who they consider to be "expert". Much depends upon the accuracy and reliability of such individuals, and it should be said that it is rare for examination boards to challenge such judgements.

However, recent research suggests that it may be unwise to leave "expert" judgements unchallenged. In a recent paper, I present results (Alderson, 1990) of studies of judgements in three areas: content inspection, item difficulty and pass-fail grades. I have already alluded to the first area above: in two studies, I showed that "expert" judges do not agree with each other on what is being tested on a range of reading tests. Moreover, where there was agreement among judges, this did not necessarily agree with the intentions of the test constructor, nor with what students reported of their test-taking processes. It is much more difficult than we may have thought to decide by content inspection alone what a test is testing. Yet much testing practice assumes we can make such judgements.

In a second study, I showed that test writers, experienced test scorers, and experienced teachers, were unable to agree on the difficulty of a set of items, for a given population, and were unable to predict the actual difficulty of items. This shows clearly the need for pre-testing of items, or at the very least for post-hoc adjustments in test content, after an analysis of item difficulty. Declarations of the suitability, or even unsuitability of a test for a given population are likely to be highly inaccurate.

In a third study, I investigated agreement among judges as to the suitability of cut-offs for grades in school-leaving examinations. There was considerable disagreement among judges as to what score represented a "pass ", a "credit" and a "distinction" at O Level. Interestingly, it proved possible to set cut-offs for the new examination by pooling the judgements - the result of that exercise came remarkably close to a norm-referenced percentile equating method. Nevertheless, the amount of disagreement as to what constituted an adequate performance for students is worrying; a worry that is confirmed by a recent review of standard-setting procedures by Berk, who shows the instability and variability of judgements (Berk, 1986).

What will clearly be needed in the coming years is a set of studies into the accuracy and nature of the range of judgements that language testers are required to make, in order to identify ways in which such judgements can be made reliable and also more valid. Testing depends upon judgements by "experts". We need to know how to improve these judgements, and how to guarantee their reliability and validity.

8. Traditional concerns

This question of the reliability and validity of judgements in testing brings me on to my final point, which relates to the traditional concerns of language testers and users of language tests. Are the tests valid? Are they reliable? What standards are followed to ensure reliability and validity?

Any review of research and practice in language testing reveals an ongoing concern with test validation. The past decade has indeed seen the introduction of new ways to validate tests, both statistical through Item Response Theory, Confirmatory Factor Analysis, MultiTrait, Multimethod analyses of convergent and discriminant validities, and the like, and qualitative, through introspective studies, through comparisons with "real-life" language use, through increased sensitivity to developments in language teaching and applied linguistics, increased concern for the "communicative" nature of the tests, and so on. We have also seen attempts to devise new test methods that might help us to reduce method bias, and it is increasingly commonplace to advocate at least the triangulation of test methods (ie the use of more than one test method in any test battery) in order to maximise our chances of measuring traits, not just test method abilities.

Clearly more needs to be done in the validation of the new tests we produce - this certainly applies to the so-called communicative tests like the CUEFL and the IELTS, but also more generally across the range of language tests for which we are responsible. But in addition to this, I believe that the 1990s will be a period during which there will be increasing pressure from test users and from test researchers for accountability: accountability of test quality and the meaning and interpretation of test scores. I have already mentioned in passing the proposed development of a professional association of language testing specialists. Associated with that will come a requirement that we develop a set of standards for good practice in language testing. There already exist general standards for educational and psychological testing - the APA, AERA and NCME standards. However, these do not refer specifically to language tests, nor, I believe, do they take account of, or accommodate to, the variety of test development procedures that exist around the world. The TOEFL - Cambridge Comparability Study I have referred to revealed considerable differences in approaches to test development, and to the establishing of standards for tests in the United Kingdom and the USA. Similar cross-national comparisons elsewhere would doubtless also reveal considerable differences. What we need to do is not to impose one set of standards on other systems, but to explore the advantages and disadvantages, the positives and negatives of the different traditions that might emerge from a survey of current practice, and to incorporate the positive features of current practice into a set of standards that could - should? - be followed by those who develop language tests. There already exists a well-documented and well-articulated psychometric tradition for establishing test standards, especially but not exclusively in the USA. What we now need to do is to identify the positive features of other traditions, and to explore the extent to which these are compatible or incompatible with the psychometric tradition.

Clearly this will take time, and considerable effort, and may well cause some anguish. Just as does a long-distance run. The analogy may not be entirely inappropriate, since the effort may well need stamina and determination. And the end-point may well be distant. But I believe it to be worthwhile.

To summarise: Recent research is beginning to challenge some of the basic assumptions we have made in the past 20-30 years: our judgements as experts are suspect; our insights into test content and validity are challengeable; our methods of test analysis may even be suspect. The apparent progress we think we have made - that we celebrate at conferences and seminars like this one, that we publish and publicise - may well not represent progress so much as activity, sometimes in decreasing circles.

It may at times appear, it may even during this talk have appeared as if the goal is reducing into the distance. Are we facing a mirage, in which our goals appear tantalisingly close, yet recede as we try to reach them? I believe no', but I do believe that we need patience and stamina in order to make progress. At least language testing is now mature enough, confident enough, well trained enough, to take part in the run, to begin the long distance journey. Did you know that to take part in long-distance events, at least in the United Kingdom, you have to be at least eighteen years old? Modern approaches to language testing are at least that old. We should not despair, but should identify the direction in which we want and need to move, and continue to work at it.

BIBLIOGRAPHY

Alderson, J Charles (1986) Innovations in language testing? In Portal, M (ed) Innovation in Language Testing. Windsor: NFRE-Nelson. Pages 93 - 105

Alderson, J Charles (1986) "Judgements in Language Testing". Paper presented at the 12th International Language Testing Research Colloquium, San Francisco

Alderson J Charles (1988). Innovation in Language Testing: Can The Micro Computer Help? Special Report No 1: Language Testing Update, Lancaster University.

Alderson J Charles (1990) Judgements in Language Testing. Paper presented at 12th ILTRC. San Francisco.

Alderson, J Charles and Hughes A (eds) (1981) Issue in Language Testing. ELT Documents, Number 111. London: The British Council

Alderson, J Charles and Clapham C (1989) "Applied Linguistics and Language Testing: A Case Study of the ELTS Test" Paper presented at the BAAL Conference, Lancaster, September 1989

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1985) Standards for Educational and Psychological Testing. Washington, DC: The American Psychological Association

Bachman L (1990) Fundamental Considerations in Language Testing Oxford: Oxford University Press

Berk R A (1986) "A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests". Review of Educational Research. Vol 56, No 1 pp 137-172

Buck G (1990) "Testing Listening Comprehension" Unpublished PhD thesis, University of Lancaster

Heaton B H (1988) Writing English Language Tests Second Edition. London: Longman

Henning G (1987) A Guide to Language Testing: Development, Evaluation, Research. Cambridge, Mass: Newbury House

Hughes A (1989) Testing for Language Teachers Cambridge: Cambridge University Press

Hughes A and Porter D (eds) (1983) Current Developments in Language Testing. London: Academic Press

Jones S, DesBrisay M and Paribakht T (eds) (1985) Proceedings of the 5th Annual Language Testing Colloquium. Ottawa: Carleton University

Oller, J W Jnr (ed) (1983) Issue in Language Testing Research Rowley: Newbury House

Read, J A S (ed) (1981) Directions in Language Testing. Singapore: SEAMEO Regional Language Centre/ Singapore University Press

Skehan P (1988) Language Testing, Part 1 and Part 2. State of the Art Article, Language Teaching, p211 - 221 and p 1 - 13

Skehan, P (1989) Progress in Language Testing: The 1990s Paper presented at the IATEFL Special Interest Group in Language Testing Symposium on Language Testing in the 1990s: the Communicative Legacy. Bournemouth, November, 1989

Stansfield C W (ed) (1986): Technology and Language Testing. TESOL: Washington, DC

Weir C J (1988) Communicative Language Testing. Volume 11: Exeter Linguistic Studies. Exeter: University of Exeter.