DOCUMENT RESUME

ED 365 144 FL 021 757

AUTHOR	Anivan, Sarinee, Ed.
TITLE	Current Developments in Language Testing. Anthology Series 25.
INSTITUTION	Southeast Asian Ministers of Education Organization (Singapore). Regional Language Centre.
REPORT NO	ISBN-9971-74-037-0; ISSN-0129-8895; RELCP383-91
PUB DATE	91
NOTE	271p.; Selected papers presented at the Regional Language Centre Seminar on Language Testing and Language Programme Evaluation (April 9-12, 1990). For selected individual papers, see also ED 328 072; ED 350 819, ED 322 725, ED 342 211, ED 317 073, ED 322 729.
PUB TYPE	Collected Works - General (020) -- Collected Works - Serials (022)
EDRS PRICE	MF01/PC11 Plus Postage.
DESCRIPTORS	Communicative Competence (Languages); Comparative Analysis; Computer Assisted Testing; Cues; English (Second Language); Evaluation Criteria; Instructional Materials; International Programs; Interviews; Item Response Theory; Language Aptitude; *Language Proficiency; Language Research; *Language Tests; *Second Languages; Simulation; Sociocultural Patterns; Speech Skills; Teacher Developed Materials; *Test Format; *Testing; Testing Programs; Test Validity; Trend Analysis; Writing Evaluation; Writing Exercises
IDENTIFIERS	*Oral Proficiency Testing

ABSTRACT

The selection of papers on language testing includes: "Language Testing in the 1990s: How Far Have We Come? How Much Further Have We To Go?" (J. Charles Alderson); "Current Research/Development in Language Testing" (John W. Oller, Jr.); "The Difficulties of Difficulty: Prompts in Writing Assessment" (Liz Hamp-Lyons, Sheila Prochnow); "The Validity of Writing Test Tasks" (John Read); "Affective Factors in the Assessment of Oral Interaction: Gender and Status" (Don Porter); "Authenticity in Foreign Language Testing" (Peter Doye); "Evaluating Communicative Tests" (Keith Morrow); "Materials-Based Tests: How Well Do They Work?" (Michael Milanovic); "Defining Language Ability: The Criteria for Criteria" (Geoff Brindley); "The Role of Item Response Theory in Language Test Validation" (T. F. McNamara); "The International English Language Testing System IELTS: Its Nature and Development" (D. E. Ingram); "A Comparative Analysis of Simulated and Direct Oral Proficiency Interviews" (Charles W. Stansfield); "Southeast Asian Languages Proficiency Examinations" (James Dean Brown, H. Gary Cook, Charles Lockhart, Teresita Ramos); "Continuous Assessment in the Oral Communication Class: Teacher Constructed Test" (Shanta Nair-Venugopal); and "What We Can Do with Computerized Adaptive Testing...And What We Cannot Do!" (Michel Laurier). (MSE)

CURRENT DEVELOPMENTS IN LANGUAGE TESTING

REGIONAL
LANGUAGE
CENTRE

SEAMEO

Edited by
SARINEE ANIVAN

ANTHOLOGY SERIES 25

CURRENT DEVELOPMENTS IN LANGUAGE TESTING

CURRENT DEVELOPMENTS IN LANGUAGE TESTING

Edited by Sarinee Anivan

Anthology Series 25
Published by
SEAMEO Regional Language Centre

30 Orange Grove Road
Singapore 1025
ISBN 9971-74-037-0
ISSN 0129-8895

Printed by Fong & Sons Printers Pte. Ltd.

CONTENTS

FOREWORD

Considerable interest in current language testing and the evaluation of language teaching programmes amongst both researchers and language educators was evidenced by the excellent and wide-ranging attendance at RELC's Regional Seminar on **Language Testing and Language Programme Evaluation** held from 9-12 April 1990. Many well-known personalities in the field, from within and outside the SEAMEO member countries, attended and gave papers or workshops. I am happy that RELC is able to publish selected papers that appear in this volume.

The importance of language testing is recognized by virtually all professionals in the field of language education. It is of special importance in education systems that are highly competitive, as education systems in Southeast Asia tend to be, as testing is not only an indirect stimulus to learning, but plays a crucial role in determining the success or failure of an individual's career, with direct implications for his future earning power.

Moreover, in countries where education is seen as central to socio-economic development, it is important that tests be valid and reliable. It is our belief that where validity and reliability are lacking, individuals as well as educational programmes suffer, and society at large is the loser. Thus, testing is an important tool in educational research and for programme evaluation, and may even throw light on both the nature of language proficiency and language learning.

Although the theme of the 1990 Seminar encompassed both testing and programme evaluation, it has not been possible to cover both areas in one anthology, and the present volume deals with papers on testing. Those that deal with programme evaluation will be published in a separate volume later.

The papers presented here are inevitably a selection, limitations of space having forced us to omit a number of papers. However, I am confident that the papers offered all make worthwhile contributions to the important field of language testing, and that they will be of interest to language educators both within and beyond Southeast Asia.

Earnest Lau
Director, RELC

INTRODUCTION

The hoped-for developments in language testing did not materialise during the 1980s. In spite of rapid growth in the field, many of the problems that existed a decade ago are still with us. **Alderson** in his paper in this volume sums up the past decade as one in which, "there has been much movement, a lot of it sideways and backwards and not much of it ... forwards". He describes areas where language testing has made progress, but also lists problems which are as yet intractable, and includes recommendations for a wide range of topics on which research can be targetted.

Naturally, as in all research, the approach taken and the basis of the research is crucial to the findings. **Oller** points out some of the fundamental problems of research in language testing. He argues for a greater understanding of language proficiency from the point of view of semiotic abilities and processes, as well as the different perspectives of the various people that are involved in the test making and test taking processes. The variances that show up in language proficiency tests could thus be correctly identified and their sources properly attributed and controlled. The validity of the tests would be more secure and their significance unambiguous. Language testing could then be a "proving ground" for language theories, and hence help to define the nature of language proficiency.

A number of writers in this collection discuss sources of variance in test results: prompts are discussed by **Prochnow and Hamp-Lyons**, experience by **Read**, affective factors by **Porter. Prochnow and Hamp-Lyons** report that prompts such as topic assignment and selection make a difference in grades. **Read** similarly shows that test type makes a difference in getting different performance from test takers. **Porter** argues that the gender of the interviewer is another source of variance in test results.

According to **Oller**, language learning will take place when learners can connect their text (discourse) with their own experience, and it is the congruence between tests and experience that is crucial in determining the validity of tests. So long as this congruence or connection exists, there is less need for real world experience or authenticity to be inserted into the tests. **Doyé** similarly offers the view that complete authenticity is neither possible nor desirable. He calls for some balance between authenticity and abstraction, since he believes that what we gain in validity through authenticity may result in a loss in terms of generalisability.

Broader educational traditions and educational philosophies are often reflected in the stance and balance of language tests. In the view of **Morrow**, language tests should reflect the humanistic approach, with some emphasis on authenticity. The potential positive washback effect should be an important consideration for validity.

Materials-based tests tend to better reflect the changes and variation of teaching methods. The washback effects will also be potentially positive. **Milanovic** describes a procedure through which materials-based tests can be made very reliable as well.

Alderson stresses that the slow pace of advancement in language testing is, to a large extent, due to gaps in our understanding of the structures and processes of language proficiency. As a mature discipline, it is appropriate that language testing helps to fill some of the gaps. The lack of a universally accepted yardstick is a manifestation of this gap. Users of criterion-referencing resort to rule-of-thumb measures which might be used in accordance with some teaching or learning objectives. Expert judgement may appear to be an easy approach to assessment, but there is a problem with reliability when raters tend to show variability in their judgements. **Brindley** details the short-comings of these procedures. Areas for future research are also discussed so that some of these deficiencies can be reduced.

Measurement models such as Item Response Theory (IRT) and proficiency models may complement each other to bring greater understanding of both models. **McNamara's** efforts to confirm the unidimensionality of IRT will enhance the application of IRT. He argues that some critics of the use of IRT assume for it roles which are not intended by the users. As a measurement tool, it should not be judged according to whether it will deepen our insight into the structure of language proficiency. **McNamara** firstly responds to some critics of the unidimensionality of IRT, and secondly provides some detailed analysis of data to show IRT does have unidimensional power.

Large scale as well as small, institutional, teacher- designed tests are also described by some writers in this collection.

The International English Language Testing System (IELTS) as described by **Ingram** is very much an international effort, with an ESP component for potential students in academic institutions. **Morrow** describes the Certificates of Communicative Skills in English (CCSE) tests, which test at four levels of proficiency in the four skills. For the oral skill sub-test, both IELTS and CCSE use the face-to-face interview format, which can be expensive and difficult to organise if large numbers of test-takers and large distances are involved. The problems described by other papers such as the lack of criteria, authenticity, and variability of judgement will no doubt be confronted in these tests. How well the problems are solved can only be shown through future evaluation.

Instead of using face-to-face interviews, **Stansfield** explores the potential of the Simulated Oral Proficiency Interview using the tape recorder. He shows that, to a large extent, the outcomes are comparable to those derived using the face-to-face format, but potentially have some advantages over the latter.

Most of the papers deal with tests of English; however **Brown** describes his efforts to establish comparable tests across Southeast Asian languages: Indonesian, Khmer, Tagalog, Thai and Vietnamese.

Shanta describes a programme within her institution to assess the competence of students over a long period through a series of assessment activities. Naturally, such elaborate testing projects must be limited to small numbers of students. Similarly, the use of computers at present seems to be limited to small numbers of test-takers. Alderson is optimistic about the potential usefulness of the computer as a testing tool. Laurier discusses some of his work with Computer Adaptive Testing (CAT) and describes the advantages and limitations of CAT.

For many test-takers language tests such as the IELTS and CCSE can have long-term implications for their careers, and test-takers deserve reliable and valid tests. Innovation and experimentation are, however, always necessary in order for a discipline to progress. The impact of test results on testees should always be part of the ethical consideration when writing tests.

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 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 forward 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 university 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 in 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 was happening in 1980 with what appears to be happening in 1990.

To take an international example, first. The International Language Testing 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-collequium, 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 made 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 conceined 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 know 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 doctoral 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 bodies 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 coefficient:. 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 not, 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.

CURRENT RESEARCH/DEVELOPMENT IN LANGUAGE TESTING

John W. Oller. Jr

INTRODUCTION

Without question, the most important item on the present agenda for language testing research and development is a more adequate theoretical perspective on what language proficiency is and what sources of variance contribute to its definition in any given test situation. Perhaps the least developed idea with reference to the research has been the differentiation of sources of variance that are bound to contribute to observed differences in measures of language proficiency in different test situations.

Among the sources of variance that have heretofore been inadequately sorted out are those attributable to text/discourse as opposed to authors contrasted also with audience or consumers. With respect to these three positions, which may be roughly related to Peirce's categories of thirdness, firstness, and secondness respectively, several distinct dimensions of each source may be sorted out. Among the most salient variables to be taken into consideration are background knowledge, relative language ability, and motivation of author (first person) and consumer (second person) as well as the properties that can be distinguished as pertaining to the discourse/text itself. For an example or two, these several sources of variability (and others) are discussed within a Peircean perspective relative to research on cloze procedure and several other ways of investigating coherence/comprehensibility of texts/discourses vis a vis certain producers and interpreters. It is argued that impoverished theories that fail to take the three positions of firstness, secondness, and thirdness into consideration are doomed to inadequacy. Nor is research that fails to do so apt to be reasonably interpretable. Examples of experimental research projects that do and do not consider the relevant variables are discussed. Finally, some general recommendations are offered for test development and future research.

GREETING

After ten years, it is a distinct pleasure to be back in Singapore again and to attend once more an international conference at RELC on language testing. As Charles Alderson reminded us at least "a little" has happened in the interim (since the 1980 conference) and we look forward to seeing what the next decade may bring forth. We may hope that all of us who were able to attend this year will be able to come back in ten years time. We are saddened to note that Dr. Michael Canale is no longer with us, and are reminded of our own mortality.

It is a "noble undertaking", as General Ratanakoses (Minister of Education in Thailand and President of SEAMEO) told us yesterday that we are embarked upon, but a difficult one. Therefore, if we are to stay in it for the long haul, as Alderson said, we will require a certain level of "stamina". The Director of RELC, Mr. Earnest Lau and Dr. Jakub Isman, the Director of the SEAMEO Secretariat, defined very admirably at the opening of this year's seminar the scope and limits of the problems that we grapple with and their importance to the enterprise of education especially in multilingual settings. Again and again, in papers at the conference, we are reminded of the central role of language in the communication of information, the establishment and maintenance of social norms, and in the very definition of what education is all about.

A GOAL AND A PLAN

This morning I want to speak to you about current research and development in language testing. Following the recommendation to be "audience-centered", from A. Latief in one of yesterday's sessions, and also a suggestion from Adrian Palmer, I have tried wherever possible to illustrate the various theoretical and practical concerns of my own presentation from things said at the conference. My goal is to introduce a theory of semiosis (our use of the ability we have as human beings to form sensible representations) which regards language testing as a special case. Along the way I will introduce Charles Sanders Peirce [1839-1914], the American scientist, mathematician, logician, and philosopher, best known in this century, perhaps, for having been the mentor of William James and John Dewey.

A GOLDEN RULE FOR TESTERS

In fact, having mentioned Peirce, I am reminded of something he wrote about being audience-centered. By the end of the talk, I hope you will see its relevance to all that I have to say and to the method ! have tried to employ in saying it. When he was a young man concerning the process of writing, he wrote in his private journal, "The best maxim in writing, perhaps, is really to love your reader for his own sake" (in Fisch, et al., 1982, p. 9). It is not unlike the rule laid down in the Mosaic law and re-iterated by Christ Jesus that we should love our neighbors as ourselves. It is a difficult rule, but one that every teacher in some measure must aspire to attain. Moreover, in interpreting it with reference to what I will say here today, it is convenient that it may be put in all of the grammatical persons which we might have need of in reference to a general theory of semiosis and to a more specific theory of language testing as a special case.

For instance, with respect to the first person, whether speaker or writer, it would be best for that person to try to see things from the viewpoint of the second person, the listener or reader. With reference to the second person, it would be good to see things (or to try to) from the vantage point of the first. From the view of a third person, it would be best to take both the intentions of the first and the expectations of the second into consideration. And, as Ron MacKay showed so eloquently in his paper at this meeting, even evaluators (acting in the first person in most cases) are obliged to consider the position of "stakeholders" (in the second person position). The stakeholders are the persons who are in the position to benefit or suffer most from program evaluation. They are the persons on the scene, students, teachers, and administrators, so it follows from the generalized version of Peirce's maxim for writers (a sort of golden rule for testers) that evaluators must act as if they were the stakeholders.

Therefore, with all of the foregoing in mind, I will attempt to express what I have to say, not so much in terms of my own experience, but in terms of what we have shared as a community at this conference. May it be a sharing which will go on for many years in a broadening circle of friendships and common concerns. I suppose that our common goal in the "noble undertaking" upon which we have embarked from our different points of view converging here at RELC, is to share our successes and our quandaries in such a way that all of us may benefit and contribute to the betterment of our common cause as communicators, teachers, educators, experimentalists, theoreticians, practitioners, language testers, administrators, evaluators, and what have you.

A BROADER THEORETICAL PERSPECTIVE

It seems that our natural proclivity is to be a little bit cautious about embracing new theoretical perspectives. Therefore, it is with a certain reasonable trepidation that I approach the topic of semiotic theory. Adrian Palmer pointed out that people have hardly had time to get used to the term "pragmatics" (cf. Oller, 1970) before there comes now a new, more difficult and more abstract set of terms drawn from the semiotic theory of Charles Sanders Peirce. It is true that the term "pragmatics" has been at least partially assimilated. It has come of age over the last two decades, and theoreticians around the world now use it commonly. Some of them even gladly incorporate its ideas into grammatical theory. I am very pleased to see that at RELC in 1990 there is a course listed on "Pragmatics and Language Teaching".

Well, it was Peirce who invented the term, and as we press on with the difficult task of sinking a few pilings into solid logic in order to lay as strong a foundation as possible for our theory, it may be worthwhile to pause a moment to realize just who he was.

C. S. Peirce [1839-1914]

In addition to being the thinker who invented the basis for American pragmatism, Peirce did a great deal else. His own published writings during his 75 years, amounted to 12,000 pages of material (the equivalent of 24 books of 500 pages each). Most of this work was in the hard sciences (chemistry, physics, astronomy, geology), and in logic and mathematics. During his lifetime, however, he was hardly known as a philosopher until after 1906, and his work in grammar and semiotics would not become widely known until after his death. His followers, William James [1842-1910] and John Dewey [1859-1952], were better known during their lifetimes than Peirce himself. However, for those who have studied the three of them, there can be little doubt that his work surpassed theirs (see, for example, comments by Nagel, 1959).

Until the 1980s, Peirce was known almost exclusively through eight volumes (about 4,000 pages) published by Harvard University Press between 1931 and 1958 under the title Collected Writings of Charles S. Peirce (the first six volumes were edited by Charles Hartshorne and Paul Weiss, and volumes seven and eight by Arthur W. Burks). Only Peirce scholars with access to the Harvard archives could have known that those eight volumes represented less than a tenth of his total output.

More recently, in 1979, four volumes on mathematics appeared under the editorship of Carolyn Eisele. Peirce's work on mathematics, it is claimed, rivals and surpasses the famed Principia Mathematica by Bertrand Russell and Alfred North Whitehead. In 1982 and 1984 respectively two additional tomes of Peirce's writings have been published by Indiana University Press. The series is titled Writings of Charles S. Peirce: A Chronological Edition and is expected, when complete, to contain about twenty volumes. The first volume has been edited by Max Fisch, et al., (1982) and the second by Edward C Moore, et al., (1984). In his Preface, to the first volume (p. xi), Moore estimates that it would require an additional 80 volumes (of 500 pages each) to complete the publication of the remaining unpublished manuscripts of Peirce. This would amount to a total output of 104 volumes of 500 pages each.

Nowadays even dilettantes (such as Walker Percy a popular writer of novels) consider Peirce to have been a philosopher. In fact, he was much more. He earned his living from the hard sciences as a geologist, chemist, and engineer. His father, Benjamin Peirce, Professor of Mathematics at Harvard was widely regarded as the premier mathematician of his day, yet the work of the son by all measures seems to have surpassed that of the father (cf. Eisele, 1979). Among the better known accomplishments of Charles Sanders Peirce was a mathematical improvement in the periodic table of chemistry. He was also one of the first astronomers to correctly determine the spiral shape of the Milky Way Galaxy. He generalized Boolean algebra - a development which has played an important role in the logic of modern computing. His work in the topological problem of map-making is, some say, still unexcelled.

Ernest Nagel wrote in 1959, "There is a fair consensus among historians of ideas that Charles Sanders Peirce remains the most original, versatile, and comprehensive mind this country has yet produced" (p. 185, also cited by Moore, 1984, p. xi). Noam Chomsky, the foremost linguist and language philosopher of the twentieth century, in an interview with Mitsou Ronat in 1979, said, "The philosopher to whom I feel closest - is Charles Sanders Peirce" (p. 71). In fact, it is Peirce's theory of abduction (or hypothetical inference; see Oller, 1990) that Chomsky credits as the basis for his whole approach to the study of language.

THE CRUCIAL ROLE OF INFERENCE

Peirce himself saw abstract representation and inference as the same thing. Inference, of course, is the process of supposing something on the warrant of something else, for example, that there will be rain in Singapore because of the build-up of thunderheads all about. Peirce wrote, "Inference in general obviously supposes symbolization; and all symbolization is inference. For every symbol ... contains information. And ... all kinds of information involve inference. Inference, then, is symbolization. They are the same notions" (1865, in Fisch, 1982, p. 280). The central issue of classic pragmatism, the variety advocated by Peirce, was to investigate "the grounds of inference" (1865, in Fisch, p. 286), or, in different words, the connection of symbols and combinations of them with the world of experience. However, Peirce differed from some so-called "pragmatists" because he did not see experience as supplying any basis for inference, but rather, inference as the only possible basis for experience. In this he was encouraged by his precursor Immanuel Kant, and his position would be later buttressed by none other than Albert Einstein (see pertinent writings of Einstein in Oller, 1989).

PRAGMATIC MAPPING

Figure 1 gives a view of what I term "pragmatic mapping". It is by definition the articulate linking of text (or discourse) in a target language (or in fact any semiotic system whatever), with facts of experience known in some other manner (i.e., through a different semiotic system or systems).

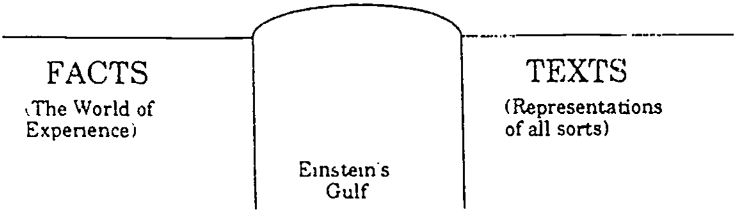


Figure 1. Pragmatic mapping.

That is, pragmatic mapping (also known as abductive reasoning), is a kind of translation process. It is a process of taking a representation in one form and interpreting it in terms of a representation in some other form. The only thing that keeps this process from being completely circular, and therefore empty, is that we really do have some valid knowledge of facts in an external world. Another point to be made is that the process of pragmatic mapping also involves risk. Or as James Pandian put it at this conference, "We talk a lot about what we don't know". Or putting the point in a slightly weaker form, we only have some of the facts most of the time and we are seeking to discover others or we may merely be speculating about them.

THE PLACE FOR SKEPTICISM

To some extent, therefore, British skepticism of the sort advocated by David Hume [1711-1776] and Bertrand Russell [1872-1970] was only partially well-founded. If there were no secure knowledge, and if all representations were always of doubtful interpretation in all circumstances (which they are not), then all representations would ultimately be meaningless, and communication and language acquisition would be impossible. However, both communication and language acquisition do in fact occur, and are in fact possible precisely because we do possess a great deal of well-equilibrated knowledge (previously established pragmatic mappings) concerning the external world--a world that is as real as the space-time continuum can be. All of this is thrashed out in detail in Oller (1989) through a collection of writings by Einstein, Peirce, James, de Saussure, Russell, Dewey, and Piaget, so that argument will not be reiterated here. Let it simply be noted that for all of its merits in pointing out the naiveness of naive realism and the positive benefits of empiricism, British skepticism failed to so much as touch the skin of classic pragmatism or the Peircean idea of abductive reasoning which forms the basis for the diagram given in Figure 1.

There are two interpretations of the figure that are of interest here. First, there is the general theory that it suggests for the comprehension of semiotic material, i.e., texts or discourse, in general, and second, there is the more specific application of it to language testing theory which we are about to develop and elaborate upon.

NECESSARY AND SUFFICIENT CONDITIONS

With respect to the first interpretation we may remark that the theory of pragmatic mapping, though entirely neglected by reviewers like Skehan (1989), offers both the necessary and sufficient conditions for language comprehension and acquisition. In order for any individual to understand any text it is necessary for that individual to articulately map it into his or her own personal experience. That is, assuming we have in mind a particular linguistic text in a certain target language, the comprehender/acquirer must determine the referents of referring noun phrases (who, what, where, and the like), the deictic significances of verb phrases (when, for how long, etc.), and in general the meanings of the text. The case is similar with the producer(s) of any given text or bit of text. All of the same connections must be established by generating surface forms in a manner that articulately corresponds to facts. If such texts are comprehended and produced (here I diverge from Krashen somewhat) over a sufficient period of time, the outcome is language acquisition. For this to occur, it figures that the individual in question must both have access to comprehensible input and must engage in comprehending it. Moreover, the learner must actively (productively) engage in the articulate linking of texts in the target language with his or her own experience. In fact, comprehension already entails this much even before any active speaking or writing ever may take place. This entails sufficient motivation in addition to opportunity. Therefore, the theory of pragmatic mapping provides both the necessary and sufficient conditions for language acquisition (whether primary or non-primary).

EINSTEIN'S GULF

Obviously, the theory requires elaboration. Before going on to a slightly elaborated diagram viewing the process in terms of a hierarchy of semiotic capacities, however, a few comments are in order concerning the middle term of Figure 1 which is referred to as "Einstein's gulf". Although it may be true that there really is an external world, and though we may know quite a lot about it (albeit practically nothing in relation to what is to be known; see the reference to Pandian above), our knowledge of the world is always in the category of being an inference. There is no knowledge of it whatever that does not involve the inferential linking of some representational form (a semiotic text of some sort) with the facts of experience. The physical world, therefore, the cosmos in all its vast extent, we do not know directly--only indirectly and inferentially through our representations of it.

The fact that physical matter should be representable at all is as Einstein put it, miraculous. He wrote of a "logically unbridgeable gulf" which "separates the world of sensory experiences from the world of concepts and propositions" (Einstein, 1944, in Oller, 1989, p. 25). This gulf poses an insurmountable barrier to any theory that would attempt to explain human intellect in a purely materialistic manner. All materialistic philosophies end in the abyss. There is for them, no logical hope whatever. It would be good to dwell on the philosophical and other implications of this, but we cannot linger here.

FACTS ARE INDEPENDENT OF SOCIAL CONSENSUS

Another point worthy of a book or two, is that what the material world is, or what any other fact in it is, i.e., what is real, in no way depends on what we may think it to be. Nor does it depend on any social consensus. Thus, in spite of the fact that our determination of what is in the material world (or what is factual concerning it), is entirely dependent on thinking and social consensus (and though both of these may be real enough for as long as they may endure), reality in general is entirely independent of any thinking or consensus. Logic requires, as shown independently by Einstein and Peirce (more elaborately by Peirce), that what is real must be independent of any human representation of it. But, we cannot develop this point further at the moment. We must press on to a more elaborate view of the pragmatic mapping process and its bearing on the concerns of language testers and program evaluators.

APPLIED TO LANGUAGE TESTING

In fact, the simplest form of the diagram, Figure 1, shows why language tests should be made so as to conform to the naturalness constraints proposed earlier (Oller, 1979, and Doye, this conference). It may go some way to explaining what Read (1982, p. 102) saw as perplexing. Every valid language test that is more than a mere working over of surface forms of a target language must require the linking of text (or discourse) with the facts of the test taker's experience. This was called the meaning constraint. The pragmatic linking, moreover, ought to take place at a reasonable speed--the time constraint. In his talk at this conference, Alderson stressed, as others have throughout, the importance of reliability and validity. It is validity that the naturalness constraints are concerned with directly.

THE SEMIOTIC HIERARCHY

Figure 2 gives a more developed view of the pragmatic mapping process. As my point of reference here at this year's RELC seminar for what follows immediately, I take N. F. Mustapha's suggestion, that we must look at the psycho-motor functions that enter into the taking of a language test.

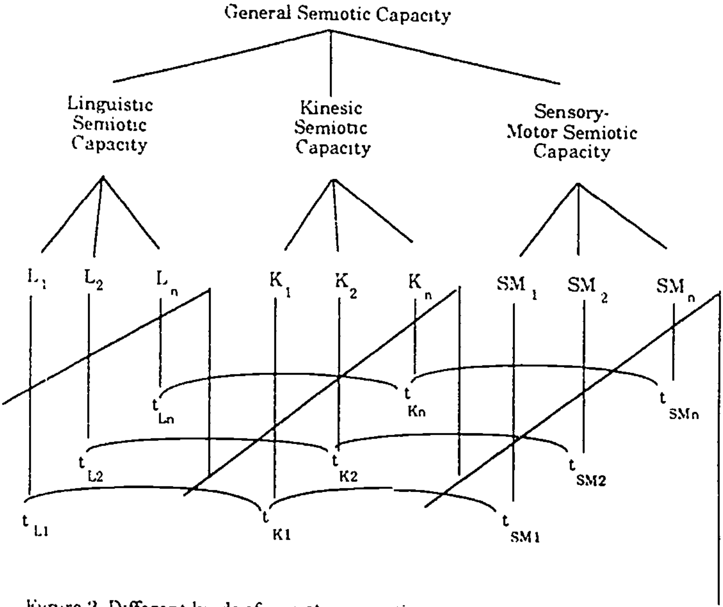

General Semiotic Capacity
Linguistic Semiotic Capacity
Kinesic Semiotic Capacity
Sensory-Motor Semiotic Capacity
L1
L2
Ln
K1
K2
Kn
SM1
SM2
SMn
tLn
tKn
tSMn
tL2
tK2
tSM2
tL1
tK1
tSM1

Figure 2. Different kinds of semiotic capacities.

The new diagram, therefore, suggests that a hierarchical organization exists. At the top of the hierarchy is what might be called general semiotic capacity. This is our ability to represent facts at the highest level of abstraction imaginable. It undergirds all the less general and more specialized capacities by which we make sense of our world. At the next level down we find at least three (perhaps there are more, but there cannot be any less) universal human capacities that are also of a representational (semiotic) sort: linguistic, kinesic, and sensory-motor. In their most abstract and general forms, each of these capacities is nonetheless distinct. Linguistic ability is the one most studied by us language testers so we may pass over it for the moment.

Kinesic Capacity. Kinesic ability pertains to our knowledge of the meanings of gestures, some aspects of which are universal and some of which are conventional and must be acquired. Smiling usually signifies friendliness, tears sadness, and so on, though gestures such as these are always ambiguous in a way that linguistic forms are not ordinarily. A smile may be the ultimate insult and tears may as well represent joy as sorrow. Sensory-motor representations are what we obtain by seeing, hearing, touching, tasting, and smelling. They include all of the visceral and other sensations of the body.

Sensory-Motor Capacity. Sensory-motor representations, as we learn from empiricism, are the starting point of all experience, experimentation, and therefore of science, and yet a little logic soon reveals that they are insufficient to determine anything by themselves (this was the valid point to be derived from the skepticism of Hume and Russell, see Oller, 1989 for elaboration). The problem with sensory-motor representations is to determine what precisely they are representations of. What do we see, hear, etc? The general logical form of the problem is a Wh-question with an indeterminate but emphatic demonstrative in it: namely, "What is *that*?" To see the indeterminacy in question, picture a scientist in a laboratory with a surprised expression on his face looking at a strange new concoction in a test-tube, or under a microscope, on a CRT, or in a mathematical formula, or wherever, and asking, "What is that?" Or imagine a person on the street or a language tester who asks the same question of any observed datum.

A gesture may help the observer determine whatever is in question. For instance, if someone points to whatever is in question or merely looks at it, this narrows down the field of possible resolutions of the demonstrative reference, but it never can adequately determine the phenomenon or object in question unless it is supported by something more abstract--namely, a conceptual or linguistic representation. With the gesture alone there is always the problem of finding out what it refers to. What precisely is pointed to or signified? In experience, gestures may serve deictic or other significant functions, but, as Peirce pointed out, gestures are always reactionally degenerate. Sensory-motor representations are also degenerate, but in a rather different way. They actually fade or dissipate over time, or even if they can be well-preserved, the physical facts themselves to which the sensory-motor impressions correspond will change and thus distort the connection between the sensory-motor representation and whatever it purports to represent.

Linguistic Capacity. Here is where language comes to the rescue. While sensory-motor representations by themselves are entirely inadequate to determine any facts about experience completely, and gestures hardly help except to bring certain significances to our attention, language affords the kind of abstract conceptual apparatus necessary to fully determine many of the facts of experience. For instance, it is only by linguistic supports that we know that today we are in Singapore, that it is Tuesday, April 9, 1990, that Singapore is an island off the southern tip of Malaysia, and west of the Philippines and north of Australia, that my name is John Oller, that Edith Hanania, Margaret Des Brisay, Liz Parkinson, Jagjeet Singh, Ron MacKay, Adrian Palmer, Kanchana Prapphal, P. W. J. Nababan, James Pandian, Tibor von Elek, and so forth, are in the audience. We know who we are, how we got to Singapore, how we plan to leave and where we would like to go back to after the meeting is over, and so forth. Our knowledge of all of these facts is dependent on linguistic representations. If any one of them were separated out from the rest, perhaps some reason could be found to doubt it, but taken as a whole, the reality suggested by our common representations of such facts is not the least bit doubtful. Anyone who pretends to think that it is doubtful is in a state of mind that argumentation and logic will not be able to cure. So we will pass on.

Particular Systems and Their Texts. Beneath the three main universal semiotic capacities identified, various particular systems are indicated. Each of these requires experience and acquisition in order to connect it to the class of texts which it defines. Each specialized semiotic system, it is asserted, supertends, or defines (in the manner of a particular grammatical system), a class of texts, or alternatively, is defined in part by the universal system that underlies it and in part by the texts that it relates to.

Relevance to Language Testing Illustrated. Now, let's see how this hierarchical model is relevant to language testing. John Read, in his very informative paper, without perhaps intending to, showed the relevance of several aspects of this model. For instance, one of the critical aspects of language use in the writing process is not merely language proficiency per se, which is represented as any given L_i, in the diagram, but is also dependent on background knowledge which may have next to nothing to do with any particular L_i. The background knowledge can only be expressed representationally as some combination of linguistic, gestural (especially indexical signs), and sensory-motor representations. It is at least obtained through such media. Perhaps in its most abstract form it is represented in purely abstract logical forms, at least part of whose structure, will be propositional in character (i.e., equilibrated relations between subjects and predicates, negations of these, and concatenations of various conjunctive and disjunctive sorts). However, knowledge which is not ultimately grounded in or related to sensory-motor contexts (i.e., sensory-motor representations) is mere superstition or pure fiction. That sort of knowledge we can know nothing of because it has no bearing on our experience.

THREE SORTS OF RESULTS PREDICTED

Looking at the pragmatic mapping process in terms of the proposed hierarchy predicts three kinds of results of immediate importance to us language testing researchers and program evaluators. Each sort of result is discussed in one way or another in papers at this conference, and it may be useful to con.ider each in turn.

(i) **Distinct Factor(s) Explained.** As John Read, Achara Wongsatorn, and Adrian Palmer showed, language proficiency can be broken into a variety of factors and, as Read argued most convincingly, language proficiency per se can properly be distinguished (at least in principle) from background knowledge. Each of the various factors (sometimes trait, sometimes skill, and sometimes method) involves different aspects of the hierarchy. For example, this can easily be demonstrated logically (and experimentally as well) with respect to the distinctness of background knowledge from language proficiency by seeing that the same knowledge can be expressed more or less equivalently in L_1, L_2, or in fact in any L_i whatever that may be known to a given user or community of users. Therefore, background knowledge is distinct from language proficiency.

(ii) **General Factor(s) Explained.** However, the hierarchical view of the theory of pragmatic mapping also shows that background knowledge and language proficiency must be inevitably interrelated. This is logically obvious from the fact that the theory (following Peirce) asserts that all comprehension and all representation is accomplished via a complex of translation processes. That is to say, if each and every semiotic representation must be understood by translating it into some other form, it follows that the various forms must have some common ground. The hypothesizing of "general semiotic capacity" at the deepest level of the hierarchy expresses this fact most perfectly, but, in fact, every node in the hierarchy suggests the interrelatedness of elements above and below that node. Hence, we have a fairly straightforward explanation for the generally high correlations betwe..n language proficiency, school achievement, IQ tests, subject matter tests, as well as the interdependency of first and second language proficiency, and many similar interactions. The general factor (more likely, factors, as John Carroll has insisted) observed in all kinds of educational or mental testing can be explained in this way.

(iii) **Non-Linearity Predicted.** The interrelatedness of elements in the hierarchy, furthermore, is bound to increase with increasing maturity and well-roundedness of experience, i.e., at higher and better integrated levels of experience. This result has been commented at this year's RELC seminar by Charles Stansfield in public discussion with Alderson (also see Oltman, Stricker, and Barrows, 1990). We see in a straightforward way why it is that as normal human beings mature, skills in all the various elements of the semiotic hierarchy are bound to mature at first at rather different rates depending on experience. This will produce, in the early stages, rather marked differences in basic skills (Figure 3) and traits (or components of language proficiency, Figure 4), just as Palmer pointed out at this seminar with reference to the sort of model that Canale and Swain, and Palmer and Bachman have argued for.

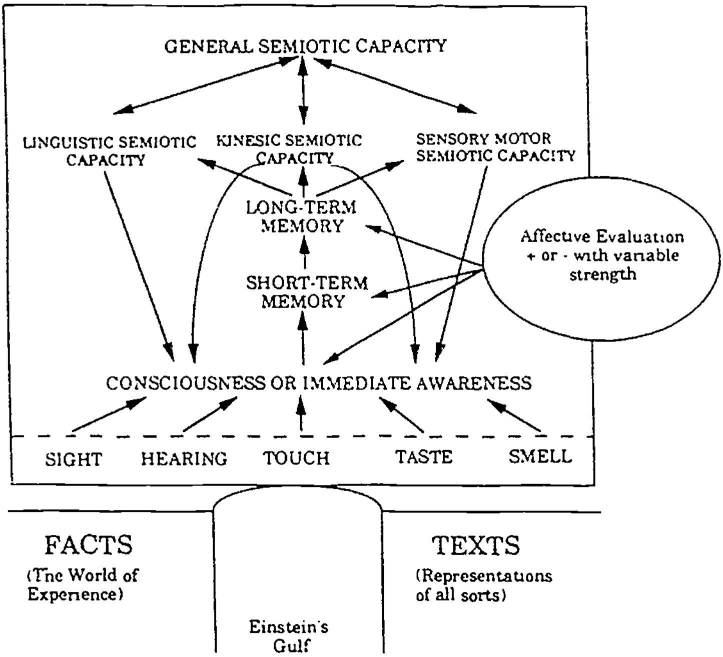

GENERAL SEMIOTIC CAPACITY
LINGUISTIC SEMIOTIC CAPACITY
KINESIC SEMIOTIC CAPACITY
SENSORY MOTOR SEMIOTIC CAPACITY
LONG-TERM MEMORY
SHORT-TERM MEMORY
Affective Evaluation + or - with variable strength
CONSCIOUSNESS OR IMMEDIATE AWARENESS
SIGHT
HEARING
TOUCH
TASTE
SMELL
FACTS
(The World of Experience)
TEXTS
(Representations of all sorts)
Einstein's Gulf

Figure 3. A modular information processing expansion of the pragmatic mapping process.

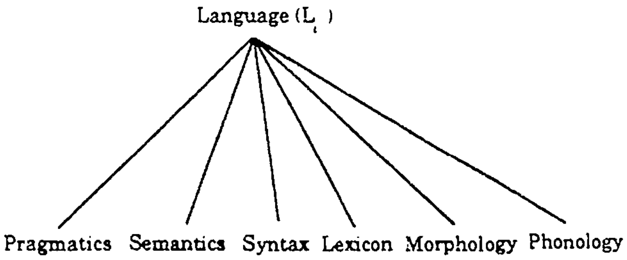


Figure 4. Language proficiency in terms of domains of grammar.

However, as more and more experience is gained, the growth will tend to fill in gaps and deficiencies such that a greater and greater degree of convergence will naturally be observed as individuals conform more and more to the semiotic norms of the mature language users of the target language community (or communities). For example, in support of this general idea, Oltman, Stricker, and Barrows (1990) write concerning the factor structure of the Test of English as a Foreign Language that "the test's dimensionality depends on the examinee's overall level of performance, with more dimensions appearing in the least proficient populations of test takers" (p. 26). In addition, it may be expected that as maturation progresses, for some individuals and groups, besides increasing standardization of the communication norms, there will be a continuing differentiation of specialized subject matter knowledge and specialized skills owing to whatever differences in experience happen to be sustained over time. For example, a person who speaks a certain target language all the time will be expected to advance in that language but not in one that is never experienced. A person who reads lots of old literary works and studies them intently is apt to develop some skills and kinds of knowledge that will not be common to all the members of a community. Or, a person who practices a certain program of sensory-motor skill, e.g., playing racquetball, may be expected to develop certain skills that a marathoner will not necessarily acquire, and so forth throughout the limitless possibilities of the hierarchy.

An Information Processing View. Another way of looking at the same basic hierarchy of semiotic capacities, still in relation to the pragmatic mapping theory, is in terms of information processing, as shown in Figure 5.

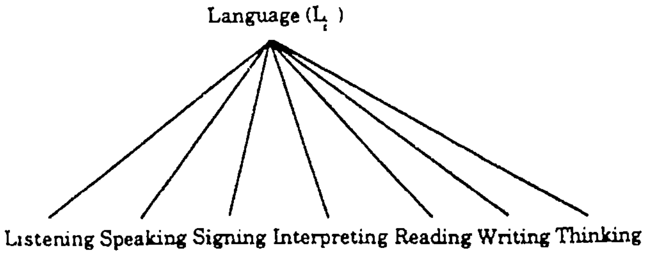


Figure 5. Language proficiency in terms of modalities of processing.

Here the general question is what sorts of internal processing go on as a language user either produces or interprets representations in relation to facts of experience. The more specific question, of interest to language testing, is how does the test taker relate the text (or discourse) of the test to the facts of his or her own experience. The general outlines of the model may be spelled out as follows. Informat'on impinges on the language user from the external world first through the senses. We might say that this is the first line of defense, and it feeds directly into consciousness or immediate awareness. At the same time consciousness is also guided by expectations coming from the various internalized grammatical systems, linguistic, kinesic, and sensory-motor. As information is processed according to these several inter-coordinated, and to some extent co-dependent expectancy systems, what is understood passes to short-term memory while whatever is not understood is filtered out as-it-were, even though it may in fact have been perceived. What is processed so as to achieve a deep level translation into a general semiotic form goes into long term memory. All the while information being processed is also evaluated affectively for its content, i.e., whether it is good (from the vantage point of the processor) or bad. In general, the distinction between a positive or negative marking, and the degree of that markedness, will determine the amount of energy devoted to the processing of the information in question. Things which are critical to the survival and well-being of the organism will tend to be marked positively in terms of affect and their absence will be regarded negatively.

Affect as Added to Cognitive Effects. The degree of importance associated with the object (in a purely abstract and general sense of the term "object") will be determined by the degree of positive or negative affect associated with it. To some extent this degree of markedness and even whether a given object of semiosis is marked positively or negatively will depend on voluntary choices made by the processor. However, there will be universal tendencies favoring survival and well-being of the organism. This means that on the positive side we will tend to find objects that human beings usually regard as survival enhancing and a complementary set of negative elements that will usually be seen as undesirable.

With respect to language processing more specifically, the consequences of affective evaluation are immense. We know of many experimental effects which show both the importance of positive and correct cognitive expectancies (these presumably from the semiotic hierarchy of capacities: linguistic, kinesic, and sensory-motor) and of positive or negative affective valuations of objects of perception, awareness, and memory. These effects are sometimes dramatic and relatively easy to illustrate. In tachistoscopic presentations of stimuli, it is well-known that contextually expected words, for instance, are easier to perceive than unexpected ones (the British psychologist John Morton comes to mind in this connection). In fact, either positive or negative expectations may be created by context which either make it easier or in fact make it harder than average to perceive a given item. These experiments carry over rather directly into the whole genre of cloze testing to which we will return shortly. However, it can be demonstrated that in addition to the effects of cognitive expectancies, affective evaluations associated with stimuli also have additional significant and important (a distinction made by James Dean Brown [1988] and alluded to by Palmer at this meeting) effects on processing. For instance, when we are hearing a conversation amid background noise and not listening, we are apt to perk up our ears so-to-speak whenever we hear our own name mentioned. It is as if the ears themselves were specially tuned for the mention of our own name. This effect and others like it, well-known to experimental psychologists are collectively known under the terms perceptual vigilance and perceptual defense. The latter phenomenon is common to the difficulty we sometimes experience in perceiving something we really don't want to see (e.g., obscenities or representations pertaining to death, and the like).

Relating all of the foregoing to language testing, I am reminded again of Read's paper of yesterday evening. As he pointed out the evidence seems to suggest that writers who are highly motivated and well-informed do better on all sorts of writing tasks. They generally write more, at a greater level of complexity, and with greater coherence. Furthermore, the graders and anyone else who takes the time to read such essays find the ones written by better motivated and better informed writers to also be that much more comprehensible. All of which leads me to the most important and final diagram for this paper, Figure 6.

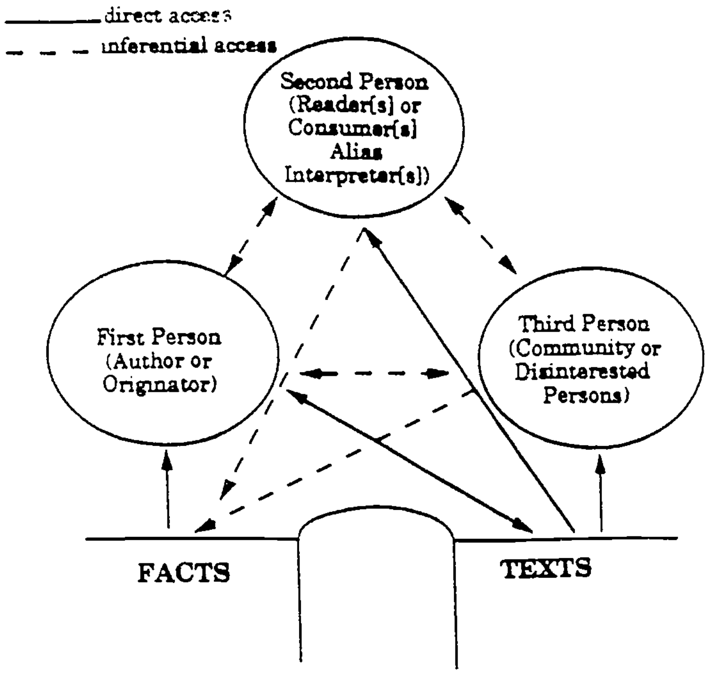

direct access
inferential access
Second Person (Reader[s] or Consumer[s] Alias Interpreter[s])
First Person (Author or Originator)
Third Person (Community or Disinterested Persons)
FACTS
TEXTS

Figure 6. The three Peircean categories as positions or perspectives of persons in reference to cloze test performances (dotted lines indicate indirect inferential connections while solid lines indicate more or less direct perceptual connections).

FIRST, SECOND, AND THIRD PERSPECTIVES

Not only is it necessary in language testing research and in program evaluation to develop a more comprehensive and better defined theoretical perspective on what semiotic capacities and processes there are, and how they interrelate with each other, but it is also, I believe, urgently necessary to differentiate the various perspectives of the persons involved in the process. The first person or producer of discourse (or text) is obviously distinct from the second person or consumer. What is not always adequately appreciated, as

Read points out in his paper at this meeting, is that variability in language tests may easily be an indiscriminant mix from both positions when only one is supposedly being tested. What is more, logically, there is a third position that is shared by the community of users (who will find the text meaningful) and the text itself. Incidentally, for those familiar with Searle's trichotomy in speech act theory (a rather narrow version of pragmatic theory), we may mention that what he calls illocutionary force (or meaning) pertains to the first position, perlocutionary force to the second, and mere locutionary force to the third.

It will be noted that the first person is really the only one who has direct access to whatever facts he or she happens to be representing the production of a particular text. Hence, the first person also has direct access to the text. At the same time the text may be accessible directly to the person to whom it is addressed, but the facts which the text represents (or purports to represent in the case of fiction) are only indirectly accessible to the second person through the representations of the first. That is, the second person must infer the intentions of the first person and the facts (whatever either of these may be). Inferences concerning those facts are based, it is hypothesized, on the sort of semiotic hierarchy previously elaborated (Figures 1-5). Similarly, a third person has direct access neither to the facts nor the intentions of the first person nor the understandings of them by the second person. All of these points must be inferred, though the text is directly accessible. The text, like the third person(s), also logically is part of the world of facts from the point of view of the third person, just as the first person and second person are part of that world. (For anyone who may have studied Peirce's thinking, the three categories differentiated here will be readily recognized as slightly corrupted, i.e., less abstract and less general, versions of his perfectly abstract and general categories of firstness, secondness, and thirdness.)

Going at these categories in a couple of different ways, I am sure that I can make clearer both what is meant by them in general and how they are relevant to the practical business of language testing. When, as language testers, we ask questions about skills and traits, as Canale and Swain (see Palmer's references) did and as Palmer and Bachman have in their several joint projects (again, see Palmer's references), we are concerned primarily in most cases with what is going on in either the first or second position. However, with some procedures attention shifts to the third position, e.g., when we use language tests to investigate characteristics of textual structure.

The point that I want to make in this next section is that unless the two other positions (beyond whichever of the three may already be in focus), and possibly a great many subtle variables within each, are controlled, it is likely that data drawn from any language testing application will be relatively meaningless. Unfortunately this is the case with far too many studies. As Palmer emphasized in his review of program evaluation with respect to theories of language acquisition and whatever sorts of proficiency may be acquired, it appears that the language teaching profession is long on methods, recipes, and hunches, and short on theories that are clear enough to put to an experimental test.

TESTING PROCEDURES AS PROVING GROUNDS

For instance, consider cloze procedure as a family of testing techniques. Between 1983 and the end of 1989 about 717 research projects of a great variety of sorts were conducted using cloze procedure in one way or another. A data search turned up 192 dissertations, 409 studies in ERIC, ε ·d 116 in the PsychLit database. At this conference there were a number of other studies that either employed or prominently referred to cloze procedure (but especially see R. S. Hidayat, S. Boonsatorn, Andrea Penaflorida, Adrian Palmer, David Nunan, and J. D. Brown). We might predict that some of the many cloze studies in recent years, not to mention the many other testing techniques, would focus on the first person position, i.e., variability attributable to the producer(s) of a text (or discourse); some on the second person position, variability attributable to the consumer(s); and some on third position, variability attributable to the text itself. Inevitably, studies o: the third position relate to factors identified with a community of language users and the sorts of texts they use.

Always a Tensional Dynamic. In fact, the interaction between a writer (or speaker) and a reader (or listener) through text (or discourse) is always a dynamic tensional arrangement that involves at least three positions simultaneously. Sometimes additional positions must be posited, but these, as Peirce showed, can always be seen as complications of the first three positions. All three of the basic positions also logically entail all of the richness of the entire semiotic hierarchy elaborated previously in this paper (Figures 1-5). Also, as John Read hinted (and as Peter Doyé stated overtly), we may move the whole theory up a level of abstraction and consider that "test-raters are different people from the test-makers, and that the way the raters interpret the task is a further source of variability in the whole process" (Read, this conference). What is not apparent in Read's statement, though I don't think he would deny it, is that the problem hinted at is completely general in language testing research and applications. All tests are susceptible to the same sort of logical criticism in terms of the sources of variability that will influence scores on them.

Congruence or ʖoodness-of-Fit as the Central Issue. In effect, the question throughout all the levels of abstraction that are imaginable, as Doyé correctly intuited though he did not say this explicitly, is whether or not the various possible positions of interlocutors (first and second positions) and texts (third), testers (first position once removed) and tests (third position once removed) interlocutors and texts, raters (first position twice removed) and testers and interlocutors and texts, etc., are in agreement. It is miraculous (as Einstein observed decades ago, see Oller, 1989) that any correspondence (i.e., representational validity) should ever be achieved between any representations and any facts, but it cannot be denied that such well-equilibrated pragmatic mappings are actually common in human experience. They are also more common than many skeptics want to admit in language testing research as well, though admittedly the testing problem is relatively (and only relatively) more complex than the basic communication problem. However, I believe that it is important to see that logically the t·vo kinds of problems are ultimately of the same class. Therefore, as testers (just as much as mere communicators) we seek convergences or "congruences" (to use the term employed by Peter Doye) between tests and what they are supposed to be tests of.

Reality and even authenticity (apart from the idea of congruence as defined within the theory of pragmatic mapping or the correspondence theory of truth which is the same thing; cf. Oller, 1990), on the other hand, are hardly worth discussing since they are so easy to achieve in their minimal forms as to be trivial and empty criteria. Contrary to a lot of flap, classrooms are real places and what takes place in them is as real as what takes place anywhere else (e.g., a train station, restaurant, ballpark, or you name it!) and to that extent tests are as real and authentic in their own right as any other superficial semiotic event. Interviews are real enough. Conversations, texts, stories, and discourse in general can be just as nonsensical and ridiculous outside the classroom (or the interview, or whatever test) as in it. Granted we should get the silliness and nonsense out of our teaching and our testing and out of the classroom (except perhaps when we are merely beir.g playful which no doubt has its place), but reality and authenticity apart from a correspondence theory of truth, or the pragmatic mapping theory outlined here, are meaningless and empty concepts.

Anything whatever that has any existence at all is ipso facto a real and authentic fact. Therefore, any test no matter how valid or invalid, reliable or unreliable, is ipso facto real ar.d, in this trivial way, authentic. The question is whether it really and authentically corresponds to facts beyond itself. But here we introduce the whole theor*/* of pragmatic mapping. We introduce all of Peirce's theory of abduction, or the elaborated correspondence theory of truth.

The test is seen as representative of something else. It is the correspondence to that something else which is really at issue. We introduce the matter of validity, truth, and goodness of fit in relation to an external world beyond the test per se. Tests, curricula, classrooms, teachers and teaching are all real enough, the problem is to authenticate or validate them with reference to what they purport to represent.

With reference to that correspondence issue, without going into any more detail than is necessary to the basic principles at stake let me refer to a few studies that show the profound differences across the several pragmatic perspectives described in Figure 6. Then I will reach my conclusion concerning all of the foregoing and hopefully justify in the minds of participants in the conference and other readers of the paper the work that has gone into building up the entire semiotic theory in the first place. There are many examples of studies focussing on the first position, though it is the least commonly studied position with cloze procedure. A dramatically clear example is a family of studies employing cloze procedure to discriminate speech samples drawn from normals from samples drawn from psychotics.

The First Person in Focus. When the first person is in focus, variability is attributable to the author (or speaker) of the text (or discourse) on which the cloze test is based. In one such study, Maher, Manschreck, Weinstein, Schneyer, and Okunieff (1988; and see their references), the third position was partially controlled by setting a task where the subjects described Breughel's "The Wedding Feast". Then cloze tests were made by replacing every fifth word with a standard blank. Paid volunteers (n = 10), then, were asked to "rate" (i.e., fill in the blanks on the various) speech samples with a minimum of two raters per sample. The assumption here being that the second position variability will be negligible. (In fact, this assumption will turn out to be wrong in this case just as it so often is in others). Results then were pooled across raters and the various authorial groups were contrasted. In fact, some discrimination did appear between different samples of speech, but (and this is the critical point to our theory), the researchers realized rather late that the second position involved variables that might drastically affect the outcomes.

A follow up study in fact aimed to test whether more educated "raters" (i.e., the paid volunteers who filled in the cloze tests) might be better at guessing all kinds of missing items and therefore might produce a ceiling effect. In such a case any differences between the speech samples of normals and psychotics would be run together at the top of the scale and thereby washed out. Indeed the follow up confirmed this expectation and it was concluded that less educated (and probably, therefore, less proficient) "raters" would generally produce greater discrimination among normal and psychotic speech samples. In addition to demonstrating that cloze procedure is sensitive to differences in the first position for psychotics and normals, this study (albeit unintentionally) showed how the procedure has to be tuned to the right level of difficulty for "raters" (i.e., persons in the second position) in order to get results. Another alternative would have been to adjust the level of difficulty of the task performed by the normals and psychotics thereby producing more complex passages (in the third position) to be cloze-rated.

Another pair of studies that focussed on first position variability with cloze procedure sought to differentiate plagiarists from students who did their own work in introductory psychology classes. In their first experiment (E1), Standing and Gorassini (1986) showed that students received higher scores on cloze passages over their own work (on an assigned topic) than over someone else's. Subjects were 16 undergraduates in psychology. In a follow-up with 22 cases, E2, they repeated the design but used a "plagiarized" essay on a new topic. In both cases, scores were higher for psychology students who were filling in blanks on their own work.

Clearly the researchers assumed in both E1 and E2 that they had sufficiently controlled the variability attributable to differences in the second position, i.e., that of the subject filling in the blanks on one or another cloze passage, and in the third, i.e., the text itself. The researchers assumed that the texts in E1 would be reasonably comparable since they were all written on an assigned topic. John Read's paper at this meeting shows that in many cases this assumption will probably not be correct. In fact, it seems fairly likely that a really bright plagiarist, one who knew the subject-matter well and who was highly proficient in the language at issue in the plagiarized material, might very well escape detection. Motivation of the writers, the amount of experience they may have had with the material, and other background knowledge are all uncontrolled variables.

With respect to E2, the third position is especially problematic. Depending on the level of difficulty of the text selected, it is even conceivable that it might be easier to fill in the blanks in the "plagiarist's" work (the essay from an extraneous source) than for some subjects to recall the exact word they themselves used in a particularly challenging essay. There is also a potential confounding of first and second positions in E1 and in E2. Suppose one of the subjects was particularly up at the time of writing the essay and especially depressed, tired, or down at the time of the cloze test. Is it not possible that an honest student might appear to be a plagiarist? Or vice versa? At any rate, difficulty, topic, level of abstraction, vocabulary employed, motivation, alertness, and a host of other factors that might be present at the time of writing and not at the filling in of the blanks (or vice versa) are potential confounding variables.

Nevertheless, there is reason to hold out hope that under the right conditions cloze procedure might be employed to discourage if not to identify plagiarists, and it should be obvious that countless variations on this theme, with reference to the first position, are possible.

The Second Person in Focus. As an example of a study focussing on the second position, consider Zinkhan, Locander, and Leigh (1986). They attempted to determine the relative effectiveness of advertising copy as judged by recallability. Two independent dimensions were identified: one affective, relating to how well the subjects (n = 420) liked the ad, brand, and product category, and one cognitive relating to knowledge and ability of the subjects (we may note that background knowledge and language proficiency are confounded here but not necessarily in a damaging way). Here, since the variability in advertising copy (i.e., third position) is taken to be a causal factor in getting people to remember the ad, it is allowed to vary freely. In this case, the first position effectively merges with the third, i.e., the texts to be reacted to. It is inferred then, on the basis of the performance of large numbers of measures aimed at the second position (the n of 420), what sorts of performances in writing or constructing ads are apt to be most effective in producing recall. In this instance since the number of cases in the second position is large and randomly selected, the variability in second position scores is probably legitimately employed in the inferences drawn by the researchers as reflecting the true qualitative reactions of subjects to the ads.

Many, if not most, second language applications of cloze procedure focus on some aspect of the proficiency or knowledge of the reader or test taker. Another example is the paper by R. S. Hidayat at this conference who wrote, "Reading as a communicative activity implies interaction between the reader and the text (or the writer through the text). To be able to do so a reader should contribute his knowledge to build a 'world' from information given by the text." I would modify this statement only with respect to the "world" that is supposedly "built" up by the reader (and or the writer). To a considerable extent both the writer and the reader are obligated to build up a representation (on the writer's side) and an interpretation (a representation of the writer's representation, on the reader's side) that conforms to what is already known of the actual world that reader, writer, and text are all part of (in defense of this see the papers by Peirce, Einstein, Dewey, and Piaget in Oller, 1989). In an even more important way, the reader's interpretation should conform in some degree to the writer's intended meaning, or else we could not say that any communication at all had occurred. Therefore, the reader had better aim to build just the world that the writer has in mind, not merely some "possible world" as so many theoreticians are fond of saying these days. Similarly, the writer, unless he or she is merely building up a fictional concoction had best have in mind the common world of ordinary experience. Even in the case of fiction writing, of course, this is also necessary to a very great extent, or else the fiction will become incomprehensible.

Happy to say, in the end, Hidayat's results are completely in line with the theory advocated here. They show a substantial correlation between the several tests aimed at grammar, vocabulary, and whatever general aspects of comprehension are measured by cloze. This is as we should expect, at least for reasonably advanced learner/acquirers. Witness prediction (ii) above that as language learners mature towards some standard level their various skills and components of knowledge will tend more and more to even out and thus to be highly correlated--producing general semiotic factors in correlational research. This being the case, apparently, we may conclude that the first and third positions were adequately controlled in Hidayat's study to produce the expected outcome in the second position.

In addition, relative to observed general factors in language testing research, recall (or refer to) the high correlations reported by Stansfield at this conference. His results are doubly confirmatory of the expected convergence of factors in the second position for relatively advanced learners (see prediction ii above) because, for one, he used a pair of rather distinct oral testing procedures, and for two, he did it with five replications using distinct language groups. In Stansfield's case, the oral tests, an Oral Proficiency Interview (OPI) and a Simulated Oral Proficiency Interview (SOPI), are themselves aimed at measuring variability in the performance of language users as respondents to the interview situation--i.e., as takers of the test regarded as if in second position. Though subjects are supposed to act as if they were in first position, since the interview is really under the control of the test writer (SOPI) or interviewer (OPI), subjects are really reactants and therefore are seen from the tester's point of view as being in second position. As Stansfield observes, with an ordinary OPI standardization of the procedure depends partly on training and largely on the wits of the interviewer in responding to the output of each interviewee.

That is to say, there is plenty of potential variability attributable to the first position. With the SOPI, variability from the first position is controlled fairly rigidly since the questions and time limits are set and the procedure is more or less completely standardized (as Stansfield pointed out). To the extent that the procedure can be quite perfectly standardized, rater focus can be directed to the variability in proficiency exhibited by interviewees (second position) via the discourse (third position) that is produced in the interview. In other words, if the first position is controlled, variability in the third position can only be the responsibility of the person in second position.

With the OPI, unlike the case of the SOPI, the interviewer (first position) variability is confounded into the discourse produced (third position). Therefore, it is all the more remarkable when the SOPI and OPI are shown to correlate at such high levels (above .90 in most cases). What this suggests is that skilled interviewers can to some extent factor their own proficiency out of the picture in an OPI situation. Nevertheless, cautions from Ross and Berwick (at this conference) and Bachman (1988) are not to be lightly set aside. In many interview situations, undesirable variability stemming from the first position (the interviewer or test designer) may contaminate the variability of interest in the second position. This caveat applies in spades to variability with respect to particular individuals interviewed though less so as the number of interviewees is increased. To avoid undesirable contamination from the first position, the interviewer (or test writer) must correctly judge the interests and abilities of the interviewee in each case so as not to place unnecessary stumbling blocks in the way. Apparently this was accomplished fairly successfully on the whole (though one wonders about individual cases) in Stansfield's study or else there would be no way to account for the surprisingly strong correlations between OPI and SOPI.

The Third Position in Focus. For a last case, consider a study by Henk, Helfeldt, and Rinehart (1985) of the third position. The aim of the study was to determine the relative sensitivity of cloze items to information ranging across sentence boundaries. Only 25 subjects were employed (second position) and two cloze passages (conflating variables of first and third position). The two passages (third position) were presented in a normal order and in a scrambled version (along the lines of Chihara, et al., 1977, and Chavez-Oller, et al., 1985). The relevant contrast would be between item scores in the sequential versus scrambled conditions. Provided the items are really the same and the texts are not different in other respects (i.e., in terms of extraneous variability stemming from first and/or second positions, or unintentional and extraneous adjustments between the scrambled and sequential conditions in the third position).

That is, the tests must not be too easy or too difficult (first position) for the subject sample tested (second position), or, alternatively, that the subject sample does not have too little or too much knowledge (second position) concerning the content (supplied by the first position) of one or both texts, the design at least has the potential of uncovering some items (third position) that are sensitive to constraints ranging beyond sentence boundaries. But does it have the potential for turning up all possible constraints of the type? Or even a representative sampling? Hardly, and there are many uncontrolled variables that fall to the first and second positions that may contaminate the outcome or prevent legitimate contrasts between the sequential and scrambled conditions from showing up even if they are really there.

In spite of this, the researchers conclude that cloze items don't do much in the way of measuring intersentential constraints. It does not seem to trouble them that this amounts to implying that they have proved that such items are either extremely rare or do not exist at all anywhere in the infinitude of possible texts. This comes near to claiming a proof of the theoretically completely general null hypothesis--that no contrast exists anywhere because none was observed here. This is never a legitimate research conclusion. Anyone can see the difficulty of the line of reasoning if we transform it into an analogous syllogism presented in an inductive order:

> Specific case, first minor premise: I found no gold in California.
> Specific case, second minor premise: I searched in two (or n) places (in California).
> General rule, or conclusion: There is no gold in California.

Anyone can see that any specific case of a similar form will be insufficient to prove any general rule of a similar form. This is not a mere question of statistics, it is a question of a much deeper and more basic form of logic.

CONCLUSION

Therefore, for reasons made clear with each of the several examples with respect to each of the three perspectives discussed, for language testing research and development to be optimally interpretable, care must be taken by researchers to control the variables of whichever of the two positions are not in focus in a particular application of any given test. In the end, in response to Jagjeet Singh (of the International Islamic University in Selangor, Malaysia) who commented that she'd have liked to get more from the lecture version of this paper than she felt she received, I have two things to say. First, that I am glad she said she wanted to receive more and flattered that "the time", as she said, "seemed to fly by" during the oral presentation (I had fun too!), and second, I hope that in years to come as she and other participants reflect on the presentation and the written version they will agree that there was even more to be enjoyed, reflected upon, understood, applied, and grateful for than they were able to understand on first pass. As Alderson correctly insists in his abstract, the study of language tests and their validity "cannot proceed in isolation from developments in language education more generally" (apropos of which, also see

Oller and Perkins, 1978, and Oller, in press). In fact, in order to proceed at all, I am confident that we will have to consider a broader range of both theory and research than has been common up till now.

REFERENCES

Bachman, Lyle. 1988. Problems in examining the validity of the ACTFL oral proficiency interview. Studies in Second Language Acquisition 10:2. 149-164.

Brown, James Dean. 1988. Understanding Research in Second Language Learning: A Teacher's guide to statistics and research design. New York: Cambridge University.

Burks, Arthur W., 1958. Collected Writings of Charles S. Peirce, Volumes VII and VIII. Cambridge, Massachusetts: Harvard University.

Chavez-Oller, Mary Anne, Tetsuro Chihara, Kelley A. Weaver, and John W. Oller, Jr. 1985. When are cloze items sensitive to constraints across sentences? Language Learning 35:2. 181-206.

Chihara, Tetsuro, John W. Oller, Jr., Kelley A. Weaver, and Mary Anne Chavez-Oller. 1977. Are cloze items sensitive to constraints across sentence boundaries? Language Learning 27. 63-73.

Chomsky, Noam A. 1979. Language and Responsibility: Based on Conversations with Mitsou Ronat. New York: Pantheon.

Eisele, Carolyn, ed. 1979. Studies in the Scientific and Mathematical Philosophy of Charles S. Peirce. The Hague: Mouton.

Fisch, Max, et al., eds., 1982. Writings of Charles S. Peirce: A Chronological Edition. Volume 1. Indianapolis: Indiana University.

Hanshorne, Charles and Paul Weiss. 1931-1935. Collected Papers of Charles Sanders Peirce. Volumes 1-6. Cambridge, Massachusetts: Belknapp of Harvard University.

Henk, William A. John P,. Helfeldt, and Steven D. Rinehart. 1985. A metacognitive approach to estimating intersentential integration in cloze tests. In National Reading Conference Yearbook 34. 213-218.

Kamil, Michael L., Margaret Smith-Burke, and Flora Rodriguez-Brown. 1986. The sensitivity of cloze to intersentential integration of information in Spanish bilingual populations. In National Reading Conference Yearbook 35. 334-338.

Maher, Brendan A., Theo C. Manschreck, Cecily C. Weinstein, Margaret L. Schneyer, and Rhoda Okunieff. 1988. Cloze analysis in schizophrenic speech: scoring method and rater's education. Perceptual and Motor Skills 67:3. 911-918.

Moore, Edward C., et al., eds., 1984. Writings of Charles S. Peirce: A Chronological Edition. Volume 2. Indianapolis: Indiana University.

Nagel, Ernest. 1959. Charles Sanders Peirce: a prodigious but little known American philosopher. Scientific American 200. 185-192.

Oller, John W., Jr. 1970. Transformational theory and pragmatics. Modern Language Journal 54. 504-507.

Oller, John W., Jr. 1979. Language Tests at School: A Pragmatic Approach. London: Longman.

Oller, John W., Jr. 1989. Language and Experience: Classic Pragmatism. Lanham, Maryland: University Press of America.

Oller, John W., Jr. 1990. Semiotic theory and language acquisition Invited paper presented at the Forty-first Annual Georgetown Round Table on Languages and Linguistics, Washington, D.C.

Oller, John W., Jr. and Kyle Perkins. 1978. Language in Education: Testing the Tests. London: Longman.

Oltman, Philip K.. Lawrence J. Stricker, and Thomas S. Barrows. 1990. Analyzing test structure by multidimensional scaling. Journal of Applied Psychology 75:1. 21-27.

Read, John A. S. 1982. Review of Language Tests at School.

John W. Oller, Jr. London: Longman, 1979. RELC Journal 13:1. 100-107.

Schumann, John. 1983. Art and science in second language acquisition research. In Language Learning, Special Issue: An Epistemology for the Language Sciences, ed. by Alexander Guiora, volume 33.5. 49-76.

Skehan, Peter. 1989. Individual Differences in Second Language Leaming. New York: Edward Amold.

Standing, Lionel G. and Donald Gorassini. 1986. An evaluation of the cloze procedure as a test for plagiarism. Teaching of Psychology 13:3. 130-132.

Zinkhan, George M, William B. Locander, and James H Leigh. 1986. Dimensional relationships of aided recall and recognition. Journal of Advertising 15:1. 38-46.

THE DIFFICULTIES OF DIFFICULTY: PROMPTS IN WRITING ASSESSMENT

Liz Hamp-Lyons and Sheila Prochnow

INTRODUCTION

In the field of writing assessment, a growing educational industry not only in the United States but also worldwide, it is often claimed that the "prompt", the question or stimulus to which the student must write a response, is a key variable. Maintaining consistent and accurate judgments of writing quality, it is argued, requires prompts which are of parallel difficulty. There are two problems with this. First, a survey of the writing assessment literature, in both L1 (Benton and Blohm, 1986; Brossell, 1983; Brossell and Ash, 1984; Crowhurst and Piche, 1979; Freedman, 1983; Hoetker and Brossell, 1986, 1989; Pollitt and Hutchinson, 1987; Quellmalz et al, 1982; Ruth and Murphy, 1988; Smith et al, 1985) and L2 (Carlson et al, 1985; Carlson and Bridgeman, 1986; Chiste and O'Shea, 1988; Cummings, 1989; Hirokawa and Swales, 1986; Park, 1988; Reid, 1989 (in press); Spaan, 1989; Tedick, 1989; Hamp-Lyons, 1990), reveals conflicting evidence and opinions on this. Second (and probably causally prior), we do not yet have tools which enable us to give good answers to the questions of how difficult tasks on writing tests are (Pollitt and Hutchinson, 1985). Classical statistical methods have typically been used, but are unable to provide sufficiently detailed information about the complex interactions and behaviors that underlie writing ability (Hamp-Lyons, 1987). Both g-theory (Bachman, 1990) and item response theory (Davidson, in press) offer more potential but require either or both costly software and statistical expertise typically not available even in moderate-sized testing agencies, and certainly not to most schools-based writing assessment programs.

An entirely different direction in education research at the moment, however, is toward the use of judgments, attitude surveys, experiential data such as verbal protocols, and a generally humanistic orientation. Looking in such a direction we see that language teachers and essay scorers often feel quite strongly that they can judge how difficult or easy a specific writing test prompt is, and are frequently heard to say that certain prompts are problematic because they are easier or harder than others. This study attempts to treat such observations and judgments as data, looking at the evidence for teachers' and raters' claims. If such claims are borne out, judgments could be of important help in establishing prompt difficulty prior to large-scale prompt piloting, and reducing the problematic need to discard many prompts because of failure at the pilot stage.

II. BACKGROUND

The MELAB, a test of English language proficiency similar to the TOEFL but containing a direct writing component, is developed by the Testing Division of the University of Michigan's English Language Institute and administered in the US and in 120 countries and over 400 cities around the world. In addition to the writing component, the test battery includes a listening component and a grammar/cloze/vocabulary/reading component (referred to as "Part 3"). There is also an optional speaking component, consisting of an oral interview. Scores on the 3 obligatory components are averaged to obtain a final MELAB score, and both component and final scores are reported. Scores are used by college or university admissions officers and potential employers in the United States in making decisions as to whether a candidate is proficient enough to carry out academic work or professional duties in English.

The writing component of the test is a 30-minute impromptu task, for which candidates are offered a choice of two topics. Topics are brief in length, usually no more than three or four lines, and intended to be generally accessible in content and prior
assumptions to all candidates. Topic development is an ongoing activity of the Testing Division, and prompts are regularly added to and dropped from the topic pool. In preparation of each test administration, topic sets are drawn from the topic pool on a rotating basis, so as to avoid repeated use of any particular topic set at any test administration site. Currently, 32 topic sets (i.e. 64 separate topics) are being used in MELAB administrations in the US and abroad and it is these topic sets, comprising 64 separate prompts, which examined in this study.

MELAB compositions are scored by trained raters using a modified holistic scoring system and a ten-point rating scale (see Appendix 1). Each composition is read independently by two readers, and by three when the first two disagree by more than one scale point. The two closest scores are averaged to obtain a final writing score. Thus, there are 19 possible MELAB composition scores (the 10 scale points and 9 averaged score points falling in between them). Compositions from all administration sites are sent to the Testing Division, where they are scored by trained MELAB raters. Inter-rater reliability for the MELAB composition is .90.

II. METHOD

Since research to date has not defined what makes writing test topics difficult or easy, our first step toward obtaining expert judgments had to be to design a scale for rating topic difficulty. Lacking prior models to build on, we chose a simple scale of 1 to 3, without descriptions for raters to use other than 1=easy, 2=average difficulty and 3=hard. Next the scale and rating procedures were introduced to 2 trained MELAB composition readers and 2 ESL writing experts, who each used the scale to assign difficulty ratings to 64 MELAB topics (32 topic sets). The four raters' difficulty ratings were then summed for each topic, resulting in one overall difficulty rating per topic, from 4 (complete agreement on a 1=easy rating) to 12 (complete agreement on a 3-hard rating). We then compared "topic difficulty" (the sum of judgments of the difficulty of each topic) to actual writing scores obtained on those topics, using 8,497 cases taken from MELAB tests administered in the period 1985-89.

Next, we categorized the 64 prompts according to the type of writing task each represents. We began with application of the topic type categories developed by Bridgeman and Carlson (1983) for their study of university faculty topic preferences. However, judges found that of Bridgeman and Carlson's nine categories, three were not usable because there were no instances of such topic types in the dataset; further, only about half of the dataset fit in the remaining six categories. The remaining half of the topics were generally found to call either for expository or for argumentative writing. The expository/argumentative distinction is of course one which has been made in many previous studies (Rubin and Piche, 1979; Crowhurst and Piche, 1979; Mohan and Lo, 1985; Quellmalz et al, 1982; etc). Another noticeable difference between topics is that some call for the writer to take a public orientation toward the subject matter to be discussed whereas others call for a more private orientation. Similar distinctions between prompts were noted by Bridgeman and Carlson (1983), who discuss differences in their various topic types in terms of what they call "degree of personal involvement", and by Hoetker and Brossell (1989) in their study of va:iations in degree of rhetorical specification and of "stance" required of the writer.

Based on these distinctions, we created a set of 5 task type categories: (1) expository/private; (2) expository/public; (3) argumentative/private; (4) argumentative/public, and (5) combination (a topic which calls for more than one mode of discourse and/or more than one orientation; an example of such a topic might be one which calls for both exposition and argumentation, or one which calls for both a personal and public stance, or even one which calls for both modes and both orientations). Examples of the five types are shown in Appendix 2. All 64 topics were independently assigned to the category, and then the few differences in categorization were resolved through discussion. Following a commonly held assumption often found in the literature (Bridgeman and Carlson, 1983; Hoetker and Brossell, 1989), we hypothesized that some topic type categories would be judged generally more difficult than others, and that expository/private topics would, on average, be judged least difficult, and argumentative/public topics most difficult. To test this prediction, we used a two-way analysis of variance, setting topic difficulty as the dependent variable and topic type as the independent variable.

III. RESULTS and INTERPRETATIONS

Topic Difficulty

When we displayed the summed topic difficulties based on four judges' scores for each of the 64 prompts, we obtained the result shown in Table 1:

Table i
Topic Difficulty for 64 MELAB Prompts

Topic Difficulty	Topic Set	No	Topic Difficulty	Topic Set	No.
4	11	A	8	42	A
4	27	A	8	43	B
4	31	B	8	44	B
4	33	B	8	45	A
4	34	B	8	49	B
5	46	B	9	12	A
5	49	A	9	18	B
6	30	B	9	21	B
6	35	B	9	22	B
6	41	A	9	23	A
6	47	A	9	24	B
7	12	B	9	31	A
7	22	A	9	33	A
7	29	B	9	35	A
7	34	A	9	46	A
7	37	B	9	50	B
7	38	A	10	11	B
7	40	A	10	13	A
7	40	B	10	24	A
7	43	A	10	23	A
8	10	A	10	30	A
8	21	A	10	39	B
8	23	B	10	42	B
8	26	A	10	45	B
8	28	A	10	47	B
8	28	B	11	10	B
8	33	A	11	13	B
8	32	B	11	18	A
8	37	A	11	26	B
8	38	B	11	27	B
8	39	A	11	44	A
8	41	B	12	50	A

Most prompts had a difficulty score around the middle of the overall difficulty scale (i.e. 8). This is either because most prompts are moderately difficult, or, and more likely, because of the low reliability of our judges' judgments. The reliability of the prompt difficulty judgments, using Cronbach's alpha, was .55.

And here was our first difficulty, and our first piece of interesting data: it seemed that claims that easy readers and language teachers can judge prompt difficulty, while not precisely untrue, are also not precisely true, and certainly not true enough for a well-grounded statistical study. When we looked at the data to discover whether the judgments of topic difficulty could predict writing score, using a two-way analysis of variance, in which writing score was the dependent variable and topic difficulty was the dependent variable, we found that our predictions were almost exactly the reverse of what actually happen (see Table 2).

Table 2: Difficulty Judgments and Writing Scores

```
ANALYSIS OF VARIANCE OF 8.CATSCOR  N= 8583 OUT OF 8583

SOURCE              DF  SUM OF SQRS  MEAN SQR   F-STATISTIC  SIGNIF

BETWEEN              8    413.31     51.663      5.2529       .0000
WITHIN            8574    84327.     9.8352
TOTAL             8582    84740.     (RANDOM EFFECTS STATISTICS)


ETA= .0698  ETA-SQR= .0049  (VAR COMP= .46927 -1  %VAR AMONG= .47)


SUMDIFF         N     MEAN      VARIANCE    STD DEV

(4)           679   8.9455      8.4439     2.9058
(5)           113   8.9823      6.5533     2.5599
(6)           737   9.1045      9.3872     3.0638
(7)          1539   9.4048      10.579     3.2526
(8)          2325   9.4705      9.5634     3.0925
(9)          1501   9.5776      10.851     3.2941
(10)         1040   9.6519      9.1242     3.0206
(11)          577   9.7660      10.763     3.2807
(12)           72   9.4028      7.1453     2.6731

GRAND        8583   9.4394      9.8742     3.1423
```

Mean writing score increased, rather than decreased, as topic difficulty increased, except for topics in the group judged as most difficult (those whose summed rating was 12, meaning all four judges had rated them as 3=difficult) As shown in Figure 1, topic difficulty as measured by "expert" judgment is unable to explain any of the variance in MELAB writing score.

Figure 1: ANOVA

Topic Difficulty and Writing Score

```
ANALYSIS OF VARIANCE OF 8.CATSCOR  N= 8583 OUT OF 10447

     SOURCE                DF   SUM SQRS    MEAN SQR    F-STAT     SIGNIF

     REGRESSION             1   372.05      372 05      37 841       0000
     ERROR               8581   84368.      9.8320
     TOTAL               8582   84740.

     MULT R=  06626   R-SQR= .00439 SE= 3.1356

     VARIABLE       PARTIAL      COEFF       STD ERROR     T-STAT     SIGNIF

     CONSTANT                    8.5291      .15179        56.190     0.
   16.SUMDIFF        06626       .11456      .18623 -1     6.1515       0000
```

Further, while the effect of judged topic difficulty on writing score is significant (p=.000C), the magnitude of the effect is about 18 times smaller than would be expected, considering the relative lengths of the writing and topic difficulty scales. That is, since the writing scale is approximately twice as long as the topic difficulty scale (19 points vs. 11 points), we would expect, assuming "even" writing proficiency (i.e. that writing proficiency increases in steps that are all of equal width) that every 1-point increase in topic difficulty would be associated with a 2-point decrease :n writing score; instead, the coefficient for topic d'fficulty effect (.11456) indicates that a 1-point increase in topic difficulty is .ctually, on average, associated with only about a 1/10-point increase in writing score. Also, it should be noted that such an increase is of little practical consequence, since a change of less than a point in MELAB writing score would have no effect either on reported level of writing performance o: on final MELAB score.

Task Type Difficulty

We had hypothesized that when topics were categorized according to topic type, the topic type categories would vary in judged difficulty level, and that the overall difficulty level of categories would vary along two continua: "orientation" (a private/public continuum), and "response mode" (an expository/argumentative continuum) (see Figure 2).

Figure 2: Response Mode, Orientation and Topic Difficulty

Predictions

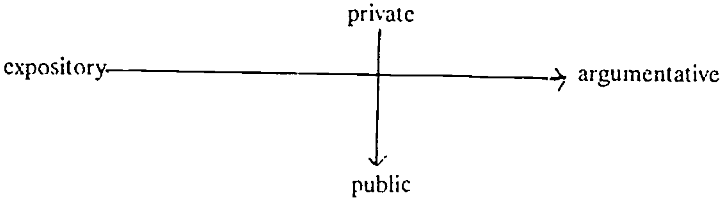


Table 3 shows the difficulty ratings for each category or "response mode":

Table 3: Response Modes and Difficulty Ratings

						Topic Category Groupings			
ExpPers		ExpPub		ArgPers		ArgPub		Comb.	
#	Diff	#	Diff	#	Diff	#	Diff	#	Diff
11A	4	40B	7	49A	5	37A	8	30B	6
27A	4	10A	8	12B	7	39A	8	34A	7
29B	4	32A	8	38A	7	43B	8	28A	8
31B	4	41B	8	38B	8	21B	9	45A	8
33B	4	49B	8	42A	8	22B	9	26B	11
34B	4	12A	9	35A	9	24B	9		
46B	5	18B	9	24A	10	31A	9		
35B	6	23A	9	29A	10	33A	9		
41A	6	10B	11	39B	10	46A	9		
47A	6	18A	11	45B	10	11B	10		
22A	7					13A	10		
37B	7					42B	10		
21A	8					13B	11		
23B	8					27B	11		
26A	8					44A	11		
28B	8					50A	12		
32B	8								
50B	9								
30A	10								
47B	10								
$\bar{x}$ diff=6.528 $\bar{x}$ wr =9.063		$\bar{x}$ diff=8.746 $\bar{x}$ wr =9.398		$\bar{x}$ diff=8.440 $\bar{x}$ wr.=9.359		$\bar{x}$ diff=8.932 $\bar{x}$ wr =9.904		$\bar{x}$ diff=7.713 $\bar{x}$ wr =9.517	

overall $\bar{x}$ diff=7.9455 overall $\bar{x}$ wr=9.4755

We conducted an ANOVA, shown in Figure 3, which showed that our predictions were correct: prompts categorized as expository/private by judges are, on average, judged easiest and those categorized as argumentative/public are judged hardest.

Figure 3: ANOVA

Topic Difficulty Judgments and Response Mode Difficulty

Judgments

```
ANALYSIS OF VARIANCE OF 16.SUMDIFF  N- 8497 OUT OF 8497

SOURCE                DF  SUM OF SQRS  MEAN SQR  F-STATISTIC SIGNIF

BETWEEN                4     8635.0     2158.8      998 42     0.
WITHIN              8492     18361      2.1622
TOTAL               8496     26996      (RANDOM EFFECTS STATISTICS)

ETA-  5656  ETA-SQR- .3199  (VAR COMP- 1.3219  %VAR AMONG- 37 94)


CATEGORY      N     MEAN      VARIANCE    STD DEV

EXPPRI      2538   6.5284      2.8666      1.6931
EXPPUB      1210   8.7463      1.6618      1.2891
ARGPRI      1543   8.4407      1 7447      1.3209
ARGPUB      2417   8.9326      1.6482      1.2838
COMBIN       789   7.7136      3.0549      1.7478

GRAND       8497   7.9854      3.1775      1.7826


 CONTRAST
 OBSERVED     PREDICTED     F-STAT     SIGNIF

 -2.0986      -0.           892.49     0.
 -2.7098      -0.           1488.0     0.
 -1.7261      -0.           603.74     0.
```

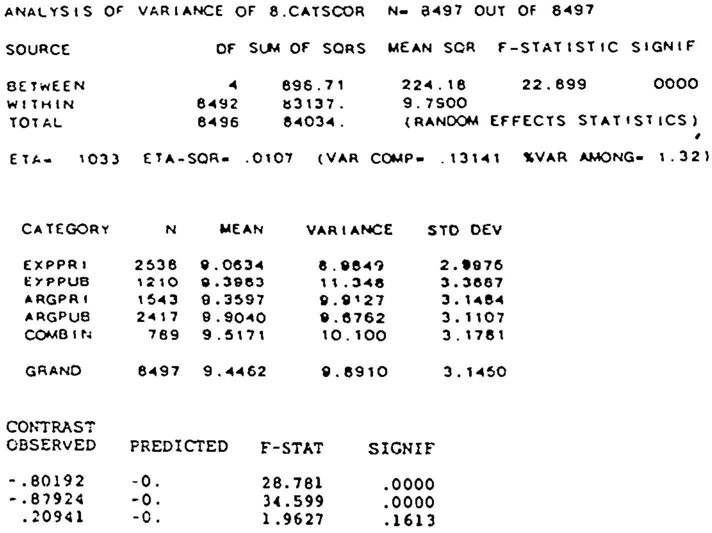
Since the two sets of judgments were made by the same judges, albeit six months apart, such a finding is to be expected.

Judgments and Writing Scores

When we looked at the relationships between our "expert" judgments of topic difficulty and task type, and compared them with writing scores, our predictions were not upheld by the data. We had hypothesized that topics in the category judged most difficult (argumentative/public) would get the lowest scores, while topics in the category judged least difficult (expository/private) would get the highest scores, with topics in the other categories falling in between. To test this hypothesis, we conducted a two-way analysis of variance, in which writing score was the dependent variable and topic type the independent variable. The results of the ANOVA, shown in Figure 4, reveal that our predictions were exactly the reverse of wha: actually happened: on average, expository/private topics are associated with the lowest writing scores and argumentative/public the highest.

Figure 4: ANOVA

Writing Performance for Prompt Categories



```
ANALYSIS OF VARIANCE OF 8.CATSCOR  N= 8497 OUT OF 8497

SOURCE              DF  SUM OF SQRS  MEAN SQR  F-STATISTIC SIGNIF

BETWEEN              4    896.71      224.18      22.899      0000
WITHIN            8492    83137.      9.7500
TOTAL             8496    84034.      (RANDOM EFFECTS STATISTICS)

ETA=  1033  ETA-SQR= .0107  (VAR COMP= .13141  %VAR AMONG= 1.32)


 CATEGORY     N     MEAN     VARIANCE   STD DEV

 EXPPRI     2538  9.0634     8.9849     2.9976
 EYPPUB     1210  9.3983     11.348     3.3687
 ARGPRI     1543  9.3597     9.9127     3.1484
 ARGPUB     2417  9.9040     9.6762     3.1107
 COMBIN      789  9.5171     10.100     3.1781

 GRAND      8497  9.4462     9.8910      3.1450

CONTRAST
OBSERVED    PREDICTED    F-STAT     SIGNIF

-.80192     -0.          28.781      .0000
-.87924     -0.          34.599      .0000
 .20941     -0.          1.9627      .1613
```

We then looked at the combined effects of topic difficulty and prompt categories, predicting that topics with the lowest difficulty ratings and of the easiest (expository/private) type would get the highest writing scores, and that topics with the highest difficulty ratings and of the hardest (argumentative/public) type would get the lowest writing scores. To test this, we again used a two-way analysis of variance, this time selecting writing score as the dependent variable and topic difficulty and topic type as the independent variables. It should be noted that in order to be able to use ANOVA for this analysis, we had to collapse the number of difficulty levels from 9 to 2, in order to eliminate a number of empty cells in the ANOVA table (i.e. some topic types had only been assigned a limited range of difficulty ratings). The results of this analysis are shown in Figure 5.

Figure 5: ANOVA

Topic Difficulty Judgments, Prompt Categories, and Writing Performance

```
diffic   type        COUNT      CELL MEANS   ST DEV
1        expri       1647       8.99454      3.01525
1        expub        215       8.27442      3.26895
1        argpri       290       9.60690      3.11886
1        argpub       431       9.97680      3.08627
1        combin       399       9.62406      3.28068
2        expri        891       9.19080      2.96185
2        expub        995       9.64121      3.34214
2        argpri      1253       9.30247      3.15372
2        argpub      1986       9.88822      3.11648
2        combin       390       9.40769      3.06995


SOURCE   SUM OF SQUARES    DF     MEAN SQUARE       F        TAIL PROB


MEAN     451627.86938       1    451627.86938   46319.54     0.0
diffic       46.57869       1        46.57869                0.0289
type        769.24715       4       192.31179                0.0
dt          357.94852       4        89.48713                0.0000
ERROR     82750.52196    8487         9.75027
```

As the ANOVA suggests and Table 4 shows clearly, our predictions were again almost the reverse of what actually happened: expository/private topics judged easiest (expri 1), as a group had the second lowest mean writing score, while argumentative/public topics judged most difficult, as a group had the second highest mean writing score.

Table 4:

Combined Effects of Topic Difficulty and Topic Type

x writing score	topic type & difficulty	
8.27442	expository/public	1
8.99454	expository/private	1
9.19080	expository/private	2
9.30247	argumentative/private	2
9.40769	combination	2
9.60690	argumentative/private	1
9.62406	combination	1
9.64121	expository/public	2
9.88822	argumentative/public	2
9.97680	argumentative/public	1

IV. DISCUSSION

Thus, patterns of relationship between topic difficulty, type and writing performance which we predicted based on commonly held assumptions were not matched by our writing score data. What we did find were unexpected but interesting patterns which should serve both to inform the item writing stage of direct writing test development, and to define questions about the effects of topic type and difficulty on writing performance which can be explored in future studies.

Several intriguing questions for further study arise from possible explanations for the patterns we did discover in our data. One possible explanation is that our judges may have misperceived what is and is not difficulty for MELAB candidates to write about. A common perceptio about writing test topics is that certain types of topics are more cognitively demanding than others, and that writers sill have more difficulty writing on these. Yet, it may be that either what judges perceive as cognitively demanding to ESL writers is in fact not, or alternately, that is not necessarily harder for ESL writers to wrie about the topics judged as more cognitively demanding while some L1 studies have concluded that personal or private topics are easier for L1 writers than impersonal or public ones, and that argumentative topics are more difficult to write on than topics calling for other discourse modes, these L1 findings do not necessarily generalize to ESL writers.

Another possible explanation for the patterns we discovered is that perhaps more competent writers choose hard topics and less competent writers choose easy topics. In fact, there is some indication in our data that this may be true. We conducted a preliminary investigation of this question, using information provided by Part 3 scores of candidates in our dataset. The Part 3 component is a 75-minute multiple choice grammar/cloze/vocabulary/reading test, for which reliability has been measured at .96(KR21). The Pearson correlation between Part 3 and writing component scores is .73, which is generally interpreted to mean that both component are measuring, to some extent, general language proficiency. We assu.ned, for our investigation of the above question, that students with a high general language proficiency (as measured by Part 3) will tend to have high writing proficiency. In our investigation we examined mean indeed been chosen by candidates with higher mean Part 3 scores. We found this to be true for 15 out of 32--nearly half--of the topic sets; thus, half of the time, general language proficiency and topic choice could account for the definite patterns of relationship we observed between judged topic difficulty, topic type and writing performance. One of these 15 sets, set 27, was used in a study by Spaan (1989), in which the same writers wrote on both topics in the set (A and B). While she found that, overall, there was not a significant difference between scores on the 2 topics, significant differences did occur for 7 subjects in her study. She attributed these differencces mostly to some subjects apparently possessing a great deal more subject matter knowledge about one topic than the other.

A further possible explanation for the relationship we observed between difficulty judgments and writing scores could be that harder topics, while perhaps more difficult to write on, push students toward better, rather than worse writing performance. This question was also explored through an investigation of topic difficulty judgments, mean Part 3 scores and mean writing scores for single topics in out dataset. We found in our dataset 3 topics whose means Part 3 scores were below average, but whose mean writing scores were average, and which were judged as "hard"(11 or 12, argumentative/public). One of these topics asked writers to argue for or against US import restrictions on Japanese cars; another asked writers to argue for or against governments treating illegal aliens differently based on their different reasons for entering; the other asked writers to argue for or against socialized medicine. The disparity between Part 3 and writing performance on these topics, coupled with the fact that they were judged as difficult, suggests that perhaps topic difficulty was an intervening variable positively influencing the writing performance of candidates who wrote on these particular topics. To thoroughly test this possibility, future studies could be conducted in which all candidates write on both topics in these sets.

A related possibility is that perhaps topic difficulty has an influence, not necessarily on actual quality of writing performance, but on raters' evaluation of that performance. That is, perhaps MELAB composition raters, consciously or subconsciously, adjust their scores to compensate for, or even reward, choice of a difficult topic. In discussions between raters involved in direct writing assessment, it is not uncom non for raters to express concern that certain topics are harder to write on than others, and that writers should therefore be given "extra credit" for having attempted a difficult topic. Whether or not these concerns translate into actual scoring adjustments is an important issue for direct writing assessment research.

V. CONCLUSION

In sum, the findings of this study provide us with information about topic difficulty judgments and writing performance without which we could effectively proceed to design and carry out research aimed at answering the above questions. In other words, we must first test our assumptions about topic

difficulty, allowing us to form valid constructs about topic difficulty, allowing us to form valid constructs about topic difficulty effect; only then can we proceed to carry out meaningful investigation of the effect of topic type and difficulty on writing performance.

REFERENCES

Bachman, Lyle. 1990. Fundamental Considerations in Language Testing. London, England: Oxford University Press.

Benton, S.L. and P.J. Blohm. 1986. Effect of question type and position on measures of conceptual elaboration in writing. Research in the Teaching of English. 20: 98-108

Bridgeman, Brent and Sybil Carlson. 1983. A Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students. TOEFL Research Report No. 15. Princeton, New Jersey: Educational Testing Service.

Brossell, Gordon. 1986. Current research and unanswered questions in writing assessment. In Greenberg, Karen S. Harvey S Weiner and Richard S. Donovan (Eds). Writing Assessment: Issues and Strategies (168-182). New York: Longman.

Brossell, Gordon. 1983. Rhetorical specification in essay examination topics. College English,45: 165-173.

Brossell, Gordon and Barbara Hoetker Ash. 1984. An experiment with the wording of essay topics. College Composition and Communication, 35: 423-425.

Carlson, Sybill and Brent Bridgeman. 1986. Testing ESL student writers. In Greenberg, Karen L., Harvey S Weiner and Richard A Donovan (Eds). Writing Assessment: Issues and Strategies(126-152). New York: Longman.

Carlson, Sybil.,Brent Bridgeman and Janet Waanders. 1985. The Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English. TOEFL Research Report 19. Princeton, New Jersey: Educational Testing Service.

Chiste, Katherine and Judith O'Shea. 1988. Patterns of Question Selection and Writing Performance of ESL Students. TESOL Quarterly, 22(4): 681-684.

Crowhurst, Marion and Gene Piche. 1979. Audience and mode of discourse effects on syntactic complexity of writing at two grade levels. Research in the Teaching of English, 13; 101-110.

Cummings,A. 1989. Writing expertise and second language proficiency. Language Learning,39(1): 81-141.

Davidson,Fred. Statistical support for reader training. In Hamp-Lyons, Liz (ed). Assessing Second Language Writing in Academic Contexts. Norwood, New Jersey: Ablex Publishing Company. In press.

Freedman, Sarah. 1983. Student characteristics and essay test writing performance. Research in the Teaching of English, '7: 313-325.

Greenberg, Karen L. 1986. The development and validation of the TOEFL writing test: a discourse of TOEFL research reports 15 and 19. TESOL Quarterly, 20(3): 531-544.

Hamp-Lyons, Liz. 1987. Testing Second Language Writing in Academic Settings. University of Edinburgh. Unpublished doctoral dissertion.

---------. 1988. The product before: task related influences on the writer. In Robinson, P (Ed). Academic Writing : Process and Product. London: Macmillan in Pauline association with the British Council.

--------. 1990. Second language writing: assessment issues. In Kroll Barbara, (Ed). Second Language Writing: Issues and Options. New York: Macmillan.

Hirokawa, Keiko and John Swales. 1986. The effects of modifying the formality level of ESL composition questions. TESOL Quarterly, 20(2): 343-345.

Hoetker, James. 1982. Essay exam topics and student writing. College Composition and Communication,33: 377-91

Hoetker, James and Gordon Brossell. 1989. The effects of systematic variations in essay topics on the writing performance of college freshman. College Composition and Communication,40(4): 414-421.

Hoetker, James and Gordon Brossell. 1986. A procedure for writing content-fair essay examination topics for large scale writing assignments. College Composition and Communication, 3, 3): 328-335.

Johns, Ann. Faculty assessment of student literacy skills: implications for ESL/EFL writing assessment. In Hamp-Lyons, Liz(Ed). Assessing Second Language Writing in Academic Contexts. Norwood, New Jersey: Ablex Publishing Company. In press.

Lunsford, Andrea 1986. The past and future of writing assessment. In Greenberg, Karen L., Harvey S. Weiner and Richard a. Donovan (Eds). Writing Assessment: Issues and Strategies. New York: Longman.

Meredith Vana H. and Paul Williams. 1984. Issues in direct writing assessment: problem identification and control. Educational Measurement: Issues and Practice. Spring 1984, 11-15, 35.

Mohan, Bernard and Winnie An Yeung Lo. 1985. Academic writing and Chinese students: transfer and developmental factors. TESOL QUARTERLY, 19(3): 515-534.

Park, Young Mok. 1988. Academic and ethnic background as factors affecting writing performance. In Purves, Alan (Ed). Writing Across Languages and Cultures: Issues in Cross Cultural Rhetoric. Newbury Park, California: Sage Publications.

Pollitt, Alistair and Carolyn Hutchinson. 1987. Calibrating graded assessment: Rasch partial credit analysis of performance in writing. Language Testing, 4(1): 72-92.

Pollitt, Alistair, Carolyn Hutchinson, Noel Gutwhistle and 1985. What Makes Exam Questions Difficult: An Analysis of 'O' Grade Questions and Answers. Research Reports for Teachers, No. 2. Edinburgh: Scottish Academic Press.

Purves, Alan, Anna Soter, Sanli Takala and A Vahapassi. 1984. Toward a domain referenced system for classifying composition assignments. Research in the Teaching of English, 18: 385-409.

Quellmalz, Edys. 1984. Toward a successful large scale writing assessment: where are we now? where do we go from here? Educational Measurement: Issues and Practice, Spring 1984: 29-32, 35.

Quellmalz, Edys, Frank Capell and Chih-Ping Chou. 1982. Effects of discourse and response mode on measurement of writing competence. Journal of Educational Measurement, 19(4): 241-258.

Reid, Joy. 1990. Responding to difference topic types: a quantitative analysis. In Kroll, Barbara (Ed). Second Language Writing Assessment: Issues and Options. New York: Macmillan.

Ruth, Leo and Sandra Murphy. 1988. Designing Writing Tasks for the Assessment of Writing. Norwood, New Jersey: Ablex Publishing Company.

________________. 1984. Designing topics for writing assessment: problems of meaning. College Composition and Communication, 35: 410-422.

Smith, W. et al. 1985. Some effects of varying the structure of a topic on college students' writing. Written Communication, 2(1): 73-89.

Spaan, Mary. 1989. Essay tests: What's in a prompt? Paper presented at 1989 TESOL convention, San Antonio, Texas, March 1989.

Tedick, Diane. 1989. Second language writing assessment: bridging the gap between theory and practice. Paper presented at the 89th Annual Convention of the American Educational Research Association, San Francisco, California, March 1989.

APPENDIX 1

MICHIGAN ENGLISH LANGUAGE ASSESSMENT BATTERY

COMPOSITION GLOBAL PROFICIENCY DESCRIPTIONS

(See reverse for composition codes)

97
Topic is richly and fully developed. Flexible use of a wide range of syntactic (sentence level) structures, and accurate morphological (word forms) control. There is a wide range of appropriately used vocabulary. Organization is appropriate and effective, and there is excellent control of connection. Spelling and punctuation appear error free.

93
Topic is fully and complexly developed. Flexible use of a wide range of syntactic structures. Morphological control is nearly always accurate. Vocabulary is broad and appropriately used. Organization is well controlled and appropriate to the material, and the writing is well connected. Spelling and punctuation errors are not distracting.

87
Topic is well developed, with acknowledgement of its complexity. Varied syntactic structures are used with some flexibility, and there is good morphological control. Vocabulary is broad and usually used appropriately. Organization is controlled and generally appropriate to the material, and there are few problems with connection. Spelling and punctuation errors are not distracting.

83
Topic is generally clearly and completely developed, with at least some acknowledgement of its complexity. Both simple and complex syntactic structures are generally adequately used; there is adequate morphological control. Vocabulary use shows some flexibility, and is usually appropriate. Organization is controlled and shows some appropriacy to the material, and connection is usually adequate. Spelling and punctuation errors are sometimes distracting.

77
Topic is developed clearly but not completely and without acknowledging its complexity. Both simple and complex syntactic structures are present; in some "77" essays these are cautiously and accurately used while in others there is more fluency and less accuracy. Morphological control is inconsistent. Vocabulary is adequate, but may sometimes be inappropriately used. Organization is generally controlled, while connection is sometimes absent or unsuccessful. Spelling and punctuation errors are sometimes distracting.

73
Topic development is present, although limited by incompleteness, lack of clarity, or lack of focus. The topic may be treated as though it has only one dimension, or only one point of view is possible. In some "73" essays both simple and complex syntactic structures are present, but with many errors; others have accurate syntax but are very restricted in the range of language attempted. Morphological control is inconsistent. Vocabulary is sometimes inadequate, and sometimes inappropriately used. Organization is partially controlled, while connection is often absent or unsuccessful. Spelling and punctuation errors are sometimes distracting.

67
Topic development is present but restricted, and often incomplete or unclear. Simple syntactic structures dominate, with many errors; complex syntactic structures, if present, are not controlled. Lacks morphological control. Narrow and simple vocabulary usually approximates meaning but is often inappropriately used. Organization, when apparent, is poorly controlled, and little or no connection is apparent. Spelling and punctuation errors are often distracting.

63
Contains little sign of topic development. Simple syntactic structures are present, but with many errors; lacks morphological control Narrow and simple vocabulary inhibits communication. There is little or no organization, and no connection apparent. Spelling and punctuation errors often cause serious interference.

57
Often extremely short; contains only fragmentary communication about the topic. There is little syntactic or morphological control. Vocabulary is highly restricted and inaccurately used. No organization or connection are apparent. Spelling is often indecipherable and punctuation is missing or appears random.

53
Extremely short, usually about 40 words or less. Communicates nothing, and is often copied directly from the prompt. There is little sign of syntactic or morphological control. Vocabulary is extremely restricted and repetitively used. There is no apparent organization or connection. Spelling is often indecipherable and punctuation is missing or appears random.

N.O.T.
N.O.T. (Not On Topic) indicates a composition written on a topic completely different from any of those assigned. It does not indicate that a writer has merely digressed from or misinterpreted a topic. N.O.T. compositions often appear prepared and memorized. They are not assigned scores or codes. 1/10/90

APPENDIX 1 (CONT'D)

MICHIGAN ENGLISH LANGUAGE ASSESSMENT BATTERY

COMPOSITION CODES

(See reverse for composition global proficiency descriptions)

NOTE: the codes are meant to indicate that a certain feature is ESPECIALLY GOOD OR BAD IN COMPARISON TO THE OVERALL LEVEL OF THE WRITING

CODE	INTERPRETATION
a	topic especially poorly or incompletely developed
b	topic especially well developed
c	organization especially inappropriate to material
d	organization especially uncontrolled
e	organization especially well controlled
f	connection especially poor
g	connection especially smooth
h	syntactic (sentence level) structures especially simple
i	syntactic stuctures especially complex
j	syntactic structures especially uncontrolled
k	syntactic structures especially controlled
l	especially poor morphological (word forms) control
m	especially good morphological control
n	vocabulary especially narrow
o	vocabulary especially broad
p	vocabulary use especially inappropriate
q	vocabulary use especially appropriate
r	spelling especially inaccurate
s	punctuation especially inaccurate
t	paragraph divisions missing or apparently random
u	handwriting illegible or nearly illegible
v	question misinterpreted or not addressed
w	reduced one score level for unusual shortness
x	other (write-in: see score report)

BEST COPY AVAILABLE

APPENDIX 2: Samples of Topic Categories

Type 1: EXPOSITORY/PRIVATE

When you go to a party, do you usually talk a lot, or prefer to listen? What does this show about your personality?

Type 2: EXPOSITORY/PUBLIC

Imagine that you are in charge of establishing the first colony on the moon. What kind of people would you choose to take with you? What qualities and skills would they have?

Type 3: ARGUMENTATIVE/PRIVATE

A good friend of yours asks for advice about whether to work and make money of whether to continue school. What advice would you give him/her?

Type 4: ARGUMENTATIVE/PUBLIC

What is you opinion of mercenary soldiers (those who are hired to fight for a country other than their own?)
Discuss.

Type 5: COMBINATION (ARGUMENTATIVE/EXPOSITORY/PUBLIC)

People who have been seriously injured can be kept alive by machines. Do you think they should be kept alive at great expense, or allowed to die? Explain your reasons.

THE VALIDITY OF WRITING TEST TASKS

John Read

INTRODUCTION

There is a long tradition in the academic world of using essays and other forms of written expression as a means of assessing student proficiency and achievement. In earlier times essay writing seemed to be quite a straightforward method of examining student performance. However, with the development of the modern science of language testing, writing has come to be considered one of the more difficult skills to test adequately because we now recognise the importance of achieving a satisfactory level of both intra-rater and inter-rater reliability in the marking of such tests. It is no longer considered acceptable to rely simply on the subjective judgement of a single teacher who has not been specifically trained or guided for the task -- although it has to be admitted that this is an idea that dies hard in the context of academic assessment in the university.

The modern concern about achieving reliability in marking has meant that relatively less attention has been paid to the other major issue in the testing of writing: how to elicit samples of writing from the students. Recent research on writing involving both native speakers and second language learners raises a number of questions about the setting of writing test tasks. The relevant research involves not only the analysis of writing tests but also more basic studies of the nature of the writing process. Some of the questions that arise are as follows:

1 To what extent is performance influenced by the amount of prior knowledge that writers have about the topic that they are asked to write about in a test?

2 Does it make a difference how the writing task is specified on the test paper?

3 Do different types of task produce significant differences in the performance of learners in a writing test?

The purpose of this paper is to explore these questions, with particular reference to the author's experience in testing English for academic purposes, and then to consider the more general issue of what constitutes a valid writing test task.

THE ROLE OF PRIOR KNOWLEDGE

One starting point in the selection of writing test tasks is a consideration of the role that knowledge of the subject matter might play in writing performance. In the case of reading, it is now widely accepted - on the basis of the research by Carrell (e.g. 1984, 1987), Johnson (1981) and others - that background knowledge is a very significant factor in the ability of second language readers to comprehend a written text. Furthermore, testing researchers such as Alderson and Urquhart (1985) and Hale (1988) have produced some evidence that a lack of relevant background knowledge can affect performance on a test of reading comprehension. Until recently, there have been few comparable studies of the role of background knowledge in second language writing, but it seems reasonable to expect that it does have a similar effect: someone is likely to write better about a familiar topic than an unfamiliar one.

Of course, in general terms this factor has long been recognised as a significant one in the testing of writing, and there are various ways in which testers have sought to minimise its effect. One approach is to give careful consideration to the choice of topic. Jacobs, et al. (1981:12-15) suggest, among other things, that the topic should be appropriate to the educational level and interests of the students; it should motivate them to communicate with the reader and should not be biased in favour of any particular sub-group among them. In other words, it should be a subject about which all potential test-takers have enough relevant information or opinions to be able to write to the best of their ability. On the other hand, it should not be too simple or predictable.

Another solution is to give the test-takers a choice of topics. In this case, it is assumed that there is a range of interests and backgrounds represented among the test-takers, and so it is hoped that all of them will find at least one that motivates them to write as best they can.

However, an alternative approach to the problem of the effect of background knowledge is to design tasks that provide the test-takers with relevant content material to work with. Thus, although differences in prior knowledge of the topic are not eliminated, the students are all provided with subject matter to use in completing the writing task, so that the focus of their efforts is not so much on generating ideas but more on expressing the ones provided in an appropriate manner.

A CLASSIFICATION OF WRITING TEST TASKS

In order to provide a basis for analysing writing test tasks according to the amount of content material that they provide, it is useful to refer to Nation's (1990) classification of language learning tasks. In this system, tasks are categorised according to the amount of preparation or guidance that the learners are given. If we adapt the classification to apply to the testing of writing, there are three task types that are relevant and they may be defined for our purposes as follows:

1 Independent tasks: The learners are set a topic and expected to write on it without any guidance.

This approach to the assessment of writing, known as the timed impromptu test, is commonly used in universities in the United States, especially in large-scale placement and proficiency tests. It assumes that all of the test-takers have background knowledge in the form of information, ideas and opinions that are relevant to the topic set. Another term which describes this type of task is "free composition".

In their simplest form, independent writing tasks can be illustrated by means of these topics, which are typical of those which are used in the composition section of the Michigan English Language Assessment Battery (MELAB):

> The role of agriculture in my country today
>
> Why young people in my country need a college education
>
> Meeting the energy needs of a modern world - problems and prospects
>
> (Quoted in Jacobs, Zingraf et al., 1981)

2 Guided tasks: The learners are provided with guidance while they are writing, in the form of a table, a graph, a picture or relevant language material.

Here we are not referring to "guided composition", in which lower proficiency learners are given language support, but rather tasks which provide content support, especially in the form of non-linear text material.

One major test which uses guided tasks of this second kind is the Test of Written English, the direct writing component of TOEFL (Test of English as a

Foreign Language). In one of the two alternating versions of this test, the candidates are presented with data in the form of a graph or a chart and are asked to write an interpretation of it (Educational Testing Service, 1989: 9). For example, a preliminary version of the test included three graphs showing changes in farming in the United States from 1940 to 1980, the task being to explain how the graphs were related and to draw conclusions from them.

3 Experience tasks: The students are given the opportunity to acquire relevant content and skills through prior experience before they undertake the writing task.

Tasks of this kind are found in major EAP tests such as the International English Language Testing Service (IELTS) and the Test in English for Educational Purposes (TEEP). In both of these tests, writing tasks are linked with tasks involving other skills to some extent in order to simulate the process of academic study. For example, in the first paper of TEEP, the candidates work with two types of input on a single topic: a lengthy written academic text and a ten-minute lecture. In addition to answering comprehension questions about each of these sources, they are required to write summaries of the information presented in each one (Associated Examining Board, 1984). The same kind of test design, where a writing task requires the synthesizing of information from readings and a lecture presented previously on the same topic, is found in the Ontario Test of ESL (OTESL) in Canada (Wesche, 1987).

Thus, the three types of task vary according to the amount and nature of the content material that is provided to the test-takers as part of the task specification. The assumption is that this may help to reduce the effects of differences in background knowledge among test-takers and, when the writing tasks are linked to earlier reading and listening tasks, may represent a better simulation of the process of academic study than simply giving a stand-alone writing test.

THE TASKS IN THE ELI WRITING TEST

In order to illustrate in practical terms the use of guided and experience tasks in the assessment of writing for academic purposes, let us look at the tasks used in a writing test developed at the English Language Institute of Victoria University. The test is administered at the end of a three-month EAP course for foreign students preparing for study at New Zealand universities, and forms part of a larger proficiency test battery. The test results provide a basis for reporting on the students' proficiency to sponsoring agencies (where applicable) and to the students themselves. They are not normally used for university admission decisions; other measures and criteria are employed for that purpose.

The test is composed of three tasks, as follows:

Task 1 (Guided)

The first task, which is modelled on one by Jordan (1980: 49), is an example of the guided type. The test-takers are given a table of information about three grammar books. For each book the table presents the title, author name(s), price, number of pages, the level of learner for whom the book is intended (basic, intermediate or advanced) and some other features, such as the availability of an accompanying workbook and the basis on which the content of the book is organised. The task is presented to the learners like this: "You go to the university bookshop and find that there are three grammar books available. Explain which one is likely to be the most suitable one for you by comparing it with the other two."

This is a guided task in the sense that the students are provided with key facts about the three grammar books to refer to as they write. Thus, the focus of their writing activity is on selecting the relevant information to use and organising the composition to support the conclusion that they have drawn about the most suitable book for them.

Task 2 (Experience)

For the second task, the test-takers are given a written text of about 600 words, which describes the process of steel-making. Together with the text, they receive a worksheet, which gives them some minimal guidance on how to take notes on the text. After a period of 25 minutes for taking notes, the texts are collected and lined writing paper is distributed. Then the students have 30 minutes to write their own account of the process, making use of the notes that they have made on the worksheet but not being able to refer to the original text.

It could be argued that this second task is another example of the guided type, in the sense that the students are provided with a reference text that provides them with content to use in their writing. However, it can also be seen as a simple kind of experience task, because the test procedure is divided into two distinct stages: first, reading and notetaking and then writing. While they are composing their text, the students can refer only indirectly to the source text through the notes that they have taken on it.

Task 3 (Experience)

The third task, like the second one, is intended to simulate part of the process of academic study. In this case, the preparation for the test begins five days beforehand, because in the week leading up to the test the students all study the topic on which the test task is based as part of their regular classwork. The topic used so far has been Food Additives. The classes spend about five hours during the week engaged in such activities as reading relevant articles, listening to mini-lectures by the teacher, taking notes, having a class debate and discussing how to organise an answer to a specific question related to the topic. However, the students do not practise the actual writing of the test task in class.
The week's activities are intended to represent a kind of mini-course on the topic of Food Additives, leading up to the test on the Friday, when the students are given an examination-type question related to the topic, to be answered within a time limit of 40 minutes. The question is not disclosed in advance either to the students or the teachers. A recent question was as follows:

> Processed foods contain additives.
> How safe is it to eat such foods?

This third task in the test is a clear example of an experience task. It provides the students with multiple opportunities during the week to learn about the topic (to acquire relevant prior knowledge, in fact), both through the class work and any individual studying they may do. Of course this does not eliminate differences in background knowledge among the students on the course. However, it is considered sufficient if the students' interest in the topic is stimulated and they learn enough to be able to write knowledgeably about it in the test.

TOWARDS A BROADER ANALYSIS OF TEST TASKS

The classification into independent, guided and experience types focuses attention on one important dimension of writing test tasks: the extent to which they provide content support for the test-takers. However, recent developments in the study and teaching of writing have highlighted a variety of other considerations that need to be taken into account, and it is to these that we now turn.

The literature on academic writing for native speakers emphasises the need to state explicitly some of the requirements that have traditionally been taken for granted. It is now more widely recognised that students need not only the ability to marshal content material effectively but also to tailor their writing for specific readers (or "audiences") and purposes. In addition, they need guidance on such matters as the amount to be written and form of text that they should produce. If such specifications are needed for native speakers of English, then the need is even greater in the case of foreign students who - as Hamp-Lyons (1988: 35) points out - often lack knowledge of the discourse and pragmatic rules that help to achieve success in an academic essay test.

At a theoretical level, Ruth and Murphy (1984) developed a model of the "writing assessment episode" involving three actors - the test-makers, the test-takers and the test-raters - and correspondingly three stages: the preparation of the task, the student's response to it and the evaluation of the student's response. The model highlights the potential for mismatch between the definition of the task as intended by the test-makers and as perceived by the students. Assuming that in a large-scale testing programme the test-raters are different people from the test-makers, the way that the raters interpret the task is a further source of variability in the whole process.

The tasks in the ELI writing test described earlier can be used to illustrate some of the problems of interpretation that arise. For example, in Task 1, the purpose of the task is not very clear. In real life, if one were making a decision about which grammar book to buy, one would normally weigh up the various considerations in one's mind or at most compose a list similar to the table in the task specification, rather than writing a prose text about it. Some of the students who took the test apparently had this difficulty and their solution was to compose their response in the form of a letter, either to a friend or to the manager of the university bookstore. Neither of these alternatives was entirely satisfactory, but the fact that some students responded in this way represents useful feedback on the adequacy of the task specification. The instinct of some of the teacher-raters was to penalize the letter-writers, on the grounds that there was nothing in the specification of the task to suggest that a letter should be written. However, one can equally argue that it was the task statement that was at fault.

Another kind of interpretation problem has arisen with the third task, when the question has been stated as follows:

> Processed foods contain additives.
> How safe is it to eat such foods?

The test-setter intended the first sentence to be an introductory statement that could be taken as given; it was the second sentence that was supposed to be the actual writing stimulus. However, in the most recent administration of the test, a number of the students took it as a proposition to be discussed and devoted the first half of their composition to it, before moving on to the question of the safety of the additives. Ruth and Murphy (1984: 417-418) noted the same phenomenon with a similar writing prompt for L1 students in the United States. Whereas the majority of the students considered the opening sentence of the prompt to be an operative part of the specification that needed to be referred to in their essays, none of the teacher-raters thought it necessary to do so.

Faced with such problems of interpretation, test-writers are in something of a dilemma. In a writing test, it is obviously desirable to minimise the amount of time that the students need to spend on reading the question or the task statement. On the other hand, as Hamp-Lyons (1988: 38) points out, a short statement may give inadequate guidance to the students on the kind of composition that they are expected to produce. This suggests that longer statements are necessary, but also that they must be composed with great care in order to reduce the possibility of misinterpretation.

The following example of a "state-of-the-art" writing task specification comes from a proposed test prepared by the U.S. Defense Language Institute for the selection of language teaching applicants:

- Assume that you have just returned from a trip and are writing a letter to a close friend. Describe a particularly memorable experience that occurred while you were traveling.

- This will be one paragraph in a longer letter to your friend. The paragraph should be about 100 words in length.

- You will be judged on the style and organization of this paragraph as well as vocabulary and grammar. Remember, the intended reader is a close fric (Herzog, 1988: 155

In terms of our classification, this is still an independent task, because it provides no content material for the test-takers to work with, but obviously it provides explicit specifications for the task in other respects.

Carlson and Bridgeman (1986: 139-142) give a useful summary of the factors to be considered in designing a stimulus for an academic writing test, based on their experience with the development of the Test of Written English. However, the most comprehensive system for classifying writing tasks is undoubtedly that established by Purves and his colleagues (Purves, Soter, Takala and Vahapassi, 1984) for the IEA Study of Written Composition in secondary schools in eighteen countries. Their system consists of fifteen dimensions: instruction, stimulus, cognitive demand, purpose, role, audience, content, rhetorical specification, tone/style, advance preparation, length, format, time, draft, and criteria. Obviously not all of these dimensions need to be specified in any individual writing stimulus, but they highlight the complexities involved, especially wh·n the test-setters and the test-takers do not share the same cultural and educational backgrounds.

DOES TYPE OF TASK MAKE A DIFFERENCE?

A more general issue related to the preparation of writing tests is whether the type of task makes a difference to the students' performance. No matter how carefully a task is specified, its validity may be limited if it does not provide a basis for making generalizations about the test-takers' writing ability. The issue is of particular relevance for a test like the Test of Written English (TWE) because, although two different task types are used, only one of them is included in any particular administration of the test. Thus, if task type is a significant variable, candidates may be advantaged or disadvantaged depending on which version of the test they take. Carlson and Bridgeman (1986) report that the pilot study of the TWE showed no significant differences in performance on the two types of task. However, Stansfield and Ross (1988) acknowledge that this is a question which requires further investigation.

In fact, Reid (1988) found evidence of differences between the TWE tasks when she analysed student scripts from a pre-operational version of the test using the Writer's Workbench, a computer text-analysis program which provided data on discourse fluency, lexical choice and the use of cohesion devices in the students' writing. The program revealed significant differences in the discourse features of the texts produced in response to the two different tasks.

There are a number of other recent research studies which provide evidence that type of task does make a difference in tests for both L1 and L2 writers. In their study of L1 secondary students in Scotland, Pollitt and Hutchinson (1987) used five different tasks. They found that the easier tasks were those in which the content and organization of the text were cued in various ways by the test stimulus. In addition, tasks that were closer to personal experience and spoken language (letter-writing or story-telling) were less demanding than more formal ones, like expressing an opinion on a controversial topic.

This variation according to formality was also present in Cumming's (1989) research on L2 writers at a Canadian university. The ratings of his three tasks differed significantly and, in particular, there was a clear distinction between the informal topic (a personal letter) and the more academic ones (an argument and a summary).

Of course, in a test of academic writing, one would normally expect that only more formal tasks would be set. However, a recent study by Tedick (1988) indicates that there is another distinction that is relevant to the setting of academic writing tests: that of general vs. specific topics. Tedick's subjects, who were graduate ESL students in the United States, wrote one essay on a general topic and another on a topic related to their own fields of study. The essays on the field-specific topic were longer, more syntactically complex and of higher overall quality (as measured by holistic ratings) than the essays on the general topic. Furthermore, the field-specific essays provided better discrimination of the three levels of ESL proficiency represented among the subjects. Tedick concluded that allowing students to make use of prior knowledge of their academic subject gave a better measure of their writing proficiency than the kind of general topic that is commonly used in writing tests.

These findings can be interpreted in relation to Douglas and Selinker's concept (1985) of "discourse domains", which are the content areas that are particularly meaningful and important for individual test-takers. Like Tedick, Douglas and Selinker argue that performance in tests of language production is affected by whether the test-takers can relate the test topic to their own interests and fields of specialization. Although these authors have looked specifically at oral production, it seems reasonable to expect that discourse domains may play a role in writing test performance as well.

There is clearly more research to be done to explore the variability in types of writing task. As Stansfield and Ross (1988) point out in a survey of research needs for the Test of Written English, there are two ways in which writing tasks can be shown to be meaningfully different. The first kind of evidence is psychometric: are there significant differences in the rankings of the test-takers when the ratings from each of the tasks are subjected to correlational analysis? In other words, are the different tasks measuring the same underlying construct? Secondly, one can use linguistic evidence to identify differences by means of syntactic and discourse analyses of the students' texts. Thus, measures of fluency, frequency of error and syntactic complexity can be obtained to reflect various aspects of the quality of the writing. Stansfield and Ross argue that, from the point of view of construct validity, the psychometric evidence is crucial: "if an empirical analysis of performance ratings on each task failed to show any significant variation between the two sets of ratings, one could claim that both were tapping the same construct, even if qualitative differences were found in the language used on each task." (op. cit.: 166) However, they acknowledge that at least some writing specialists consider that qualitative, linguistic differences are also important.

For the present, it seems prudent to include a variety of writing tasks in a writing test. It can be argued that this not only makes the assessment more reliable by producing several samples of writing from each student, but it contributes to the validity of the test by giving a broader basis for making generalizations about the student's writing ability.

CONCLUSION: Writing as Process

However, this leads to one final issue that is relevant to a consideration of the validity of writing tasks. The testing of writing inevitably focuses on the text that the test-taker produces or, in other words, the product rather than the process. Practical constraints normally mean that the students can be given only a limited amount of time for writing and therefore they must write quite fast in order to be able to produce an adequate composition. The preceding discussion of different tasks and ways of specifying them has concentrated on the issue of how to elicit the kind of texts that the test-setter wants, with little consideration of the process by which the texts will be produced and whether that reflects the way that people write in real life.

However, any contemporary discussion of writing assessment must take account of the major developments that have occurred over the last fifteen years in our understanding of writing processes. In the case of L1 writers, researchers such as Britton et al. (1975), Murray (1978), Perl (1980) and Graves (1983) have demonstrated how people compose a segment of text, pause to read and consider it, revise or replace it, plan further segments of text and so on. Murray (1978) described writing as "a process of discovery", through which writers explored their ideas and found out what they wanted to express. As Kelly (1989: 80) puts it, "the act of writing has great generative power, both in the sense of creating ideas and creating the language to express those ideas". Studies by Zamel (1983), Raimes (1987), Cumming (1989) and others have found that L2 writers exhibit very much the same strategies in composing text as L1 writers do.

There are a number of implications of this research for the assessment of writing. If the production of written text is not strictly linear but inherently recursive in nature, normally involving cyclical processes of drafting and revising, this suggests that test-takers need to be given time for thinking about what they are writing; in addition, explicit provision needs to be made for them to revise and rewrite what they have written.

The question, then, is how the writing process can be accommodated within the constraints of the test situation Of course, any test situation is different from the context of real-world language processing, but the disjunction between "natural" writing processes and the typical writing test is quite marked, particularly if we are interested in the ability of students to write essays, research reports and theses, rather than simply to perform in examination settings. Even a substantially increased time allocation for completing a test task does not alter the fact that the students are being required to write under constraints that do not normally apply to the writing process.

There are various ways in which one can reduce the effects of the time constraints. One means is to limit the demands of the writing test task by providing support as part of the task specification. The provision of content material, as is done with guided and experience tasks, is one way of reducing the complexity of the writing task and allowing the test-taker to focus on the structure and organization of the text.

Another, more radical approach which is gaining ground is to move away from a reliance on timed tests for writing assessment. This may not be possible in large-scale placement or proficiency tests for practical reasons, but there is now increasing interest in portfolio assessment (see, e.g., Katz, 1988: 196-198), which involves collecting a standard set of different types of writing completed by each student over a period of time (usually as part of a course) and then having them assessed according to agreed criteria by one or two teachers other than the class teacher. Once again, there are practical difficulties in implementing this approach, but it may be valuable as a supplementary method of assessment, especially in the case of postgraduate students whose primary academic writing activities are the preparation of theses and research papers.

Clearly, all of these considerations indicate that the validity of writing test tasks is a complex issue and one that is likely to receive increasing attention in the years to come.

REFERENCES

Alderson, J. C. and Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. Language Testing, 2, 192-204.

Associated Examining Board. (1984). The Test in English for Educational Purposes. Aldershot, England: Author.

Britton, J. et al. (1975). The development of writing abilities 11-18. London: Macmillan.

Carrell, P. L. (1984). The effects of rhetorical organization on ESL readers. TESOL Quarterly, 18, 441-469.

Carrell, P. L. (1987). Content and formal schemata in ESL reading. TESOL Quarterly, 21, 461-481.

Carlson, S. and Bridgeman, B. (1986). Testing ESL student writers. In K.L. Greenberg, H.S. Wiener and R.A. Donovan (Eds.) Writing assessment: issues and strategies. New York: Longman.

Cumming, A. (1989). Writing expertise and second-language proficiency. Language Learning, 39.1, 81-141.

Douglas D. and Selinker, L. (1985) Principles for language tests within the "discourse domains" theory of interlanguage: research, test construction and interpretation. Language Testing, 2.2, 205-226.

Educational Testing Service. (1989). Test of Written English guide. Princeton, NJ: Author.

Graves, D. (1983). Writing: children and teachers at work. London: Heinemann Educational Books.

Hale, G. A. (1988). Student major field and text content: interactive effects on reading comprehension in the Test of English as a Foreign Language. Language Testing, 5, 49-61.

Hamp-Lyons, L. (1988). The product before: Task-related influences on the writer. In P.C. Robinson (ed.) Academic writing: Process and product. (ELT Documents 129). London: Modern English Publications/British Council.

Herzog, M. (1988). Issues in writing proficiency assessment. Section 1: The governmen' scale. In P. Lowe, Jr. and C. W. Stansfield (Eds.) Second language proficiency assessment: Current issues. Englewood Cliffs, NJ: Prentice Hall Regents.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). Testing ESL composition: A practical approach. Rowley, MA: Newbury House.

Johnson, P. (1981). Effects on reading comprehension of language complexity and cultural background of a text. TESOL Quarterly, 15, 169-181.

Jordan, R. R. (1980). Academic writing course. London: Collins.

Katz, A. (1988). Issues in writing proficiency assessment. Section 2: The academic context. In P. Lowe, Jr. and C. W. Stansfield (Eds.) Second language proficiency assessment: Current issues. Englewood Cliffs, NJ: Prentice Hall Regents.

Kelly, P. (1989). Theory, research and pedagogy in ESL writing. In C.N. Candlin and T. F. McNamara (Eds.) Language, learning and community. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Murray, D. (1978). Internal revision: a process of discovery. In C. R. Cooper and L. Odell (Eds.) Research on composing: Points of departure. Urbana, Ill.: NCTE.

Nation, P. (1990). A system of tasks for language learning. In S. Anivan (Ed.) Language teaching methodology for the Nineties. Singapore: SEAMEO Regional Language Centre.

Perl, S. (1'80). Understanding composing. College Composition and Communication, 31.4, 363-369.

Pollitt, A. and Hutchinson, C. (1987) Calibrating graded assessments: Rasch partial credit analysis of performance in writing. Language Testing, 4.1, 72-92.

Purves, A. C., Soter, A., Takala, S., and Vahapassi, A. (1984). Towards a domain-referenced system for classifying composition assignments. Research in the Teaching of English, 18.4, 385-416.

Raimes, A. (1987). Language proficiency, writing ability and composing strategies: A study of ESL college writers. Language Learning, 37.3, 439-467.

Reid, J. M. (1988) Quantitative differences in English prose written by Arabic, Chinese, Spanish and English students. Unpublished Ph.D. dissertation, Colorado State University.

Ruth, L. and Murphy, S. (1984). Designing topics for writing assessment: Problems of meaning. College Composition and Communication, 35.4, 410-422.

Stansfield, C. W. and Ross, J. (1988). A long-term research agenda for the Test of Written English. Language Testing, 5, 160-186.

Tedick, D. J. (1988). The effects of topic familiarity on the writing performance of non-native writers of English at the graduate level. Unpublished Ph.D. dissertation, Ohio State University.

Wesche, M. B. (1987). Second language performance testing: The Ontario Test of ESL as an example. Language Testing, 4, 28-47.

Zamel, V. (1983). The composing processes of advanced ESL students: Six case studies. TESOL Quarterly, 17. 165-187.

AFFECTIVE FACTORS IN THE ASSESSMENT OF ORAL INTERACTION: GENDER AND STATUS

Don Porter

INTRODUCTION

1 LINGUISTIC COMPLEXITY AND LANGUAGE TESTS

In its internal structure and in its components, linguistic ability is extremely - if not infinitely - complex. Any attempt to summarize linguistic ability in the form of a description will necessarily have to consist of some form of simplification of the original complexity. Language tests are constructed on the basis of such simplifying descriptions of linguistic ability in general - what we might call linguistic 'models' - and are themselves devices for generating descriptions of the individual language user's ability in terms of the underlying model. So language tests, too, must simplify what they assess.

Sometimes the descriptions produced by a language test are in terms of numbers, eg '72%' or perhaps '59% in Writing, 72% in Reading' (although it is difficult to know what such descriptions of linguistic ability could mean, they are so abstract and relativistic); sometimes the descriptions are put in terms of verbal descriptions, eg:

> 'very little organisation of content; for the most part satisfactory cohesion; some inadequacies in vocabulary; almost no grammatical inaccuracies'
>
> (Based on criteria for Test of English for Educational Purposes: Weir, 1988)

But whatever form the description takes, the general headings under which the various aspects of the description fall are not God-given, inherent in the nature of language or linguistic ability, so much as imposed on a continuum of confusing and unruly data by language specialists. Language ability does not fall neatly into natural pre-existing categories, but has to be forced into man-made categories with varying degrees of success. A description which aims for completeness by having special headings for all the bits which do not quite fit may well end up by being more complex than the original language ability being described - and the more complex the description gets, the less our brains are able to grasp it in its entirety: the less it means to us. A truly useful description of a language ability, then, will be one which leaves a great deal out! What such a description will do will be to focus on various features which are felt to be particularly salient and important. That is to say, it will be founded on a theoretical model - one which, in the features it chooses to highlight, and in the way it relates those features one to another, attempts to capture the essence of the language ability. The questions for a test, then, are: how elaborate a model should it be based on if it is to avoid the criticism that it leaves out of account crucial features of the language ability to be measured; and on the other hand how much complexity can it afford to report before it runs the risk of being unusable?

2 Communicative Language Testing

Testers vary in whether they claim to be producing or discussing communicative competence tests, communicative performance tests, or simply - and conveniently - communicative tests, and views of what those various terms imply also vary considerably. There is no widely accepted overall model of communicative proficiency used as a basis for this approach to language testing. Nevertheless, there is in Britain at least a fair degree of working consensus about the sorts of characteristics such tests ought to have. We may cite just the following few as being fairly typical:

(a) Tests will be based on the needs (or wants) of learners. It would be unreasonable to assess a learner's ability to do through English something which he has no need or wish to do. A principle such as this suggests that the different needs of different learners may call for different types of linguistic ability at different levels of performance; in principle tests incorporating this idea will vary appropriately for each new set of needs in the number and type of abilities they assess, and in their appraisal of what constitutes a satisfactory level of performance. Results will be reported separately for ˬach ability in the form of a profile. We are thus immediately faced with a degree of test complexity at the points of test-content, assessment criteria, and report format.

(b) Tests will be based on language use in the contexts and for the purposes relevant to the learner. It is at least conceivable that any one of the linguistic ability types mentioned in the previous paragraph might be required in a number of distinct contexts crucial to the learner and for more than one distinct purpose in any given context. If varying context and purpose are seen as central features of natural communication, this suggests that particular contexts and purposes require particular deployments of linguistic abilities. Both context and purpose will then need to be suitably incorporated in tests and will represent two further dimensions of complexity.

(c) Tests will employ authentic texts, or texts which embody fundamental features of authenticity. These 'fundamental features' may well include appropriate format and appropriate length, both of which will vary with the type of text. Concerning length in particular, longer texts are said to require types of processing different from those needed for shorter texts. Text authenticity then implies yet another dimension of complexity.

These characteristic features, together with others, reflect the assumption that, in Oller's (1979) terms, language ability is not unitary, but in fact very divisible.

Tests already exist which seek to embody all these and other features of natural communication for more or less well-defined groups of learners. The challenge is great and the difficulties formidable. Bachman (1990) has criticised such tests as suffering from inadequate sampling and consequent lack of generalizability of their results: the descriptions of ability yielded by the test, it is argued, refer only to the needs, contexts, purposes, text-types, etc. covered in the test; needs, contexts, purposes, etc. are so multifarious it is not possible to sample them adequately for all test-takers, and perhaps not even for a single test-taker.

In the light of the already-existing difficulties posed for test construction, and of such criticisms, and of the need for a useful, practical test to avoid excessive complexity, we must think very carefully indeed before proposing that tests should incorporate yet another level of complexity by including information on the effects of affective factors in the descriptions which they yield.

3 Affective Factors

Affective factors are emotions and attitudes which affect our behaviour. We may distinguish between two kinds: predictable and unpredictable.

Unpredictable: Most teachers will be familiar with the kinds of affective factor which produce unpredictable and unrepresentative results in language tests, eg. a residue of anger after a family row or a mood of irresponsibility after some unexpected good news on the day. The fact that such moods may weaken concentration, or may lead in some other way to learners not reflecting in their performance the best that they are capable of, will obviously detract from the reliability of the description of abilities yielded by the test.

Clearly, if we can find ways of minimizing the effects of such unpredictable factors, we should do so. If the test is associated with a teaching programme, continuous assessment or a combination of continuous assessment with a formal test would be less likely to be affected by a single unrepresentative performance. On the other hand, if there is no associated teaching programme, and everything hangs on a single measure, we might try to eliminate from the subject matter of the test any topics which might be likely to touch on a raw nerve somewhere. For example, the Educational Testing Service carefully vets all essay topics for the Test of Written English for possible sources of unreliability, emotional associations being one such source.

Another possible route to eventual affect-free assessment might be to devise a programme of research to discover the kinds of test techniques which are least susceptible to emotional buffeting.

On the other hand, the attempt to eliminate emotional content from language tests, on whatever grounds, may be misconceived. Is it not the case that a fundamental and natural use of language is as a vehicle for messages with emotional associations? Imagine tests in which the learner is asked to react to or produce language with which he feels no personal involvement, and to which he feels no personal commitment. Would not such language be at least severely restricted in its range of content, and at most fundamentally unnatural? We are left in a dilemma: it is suggested that emotional content is a central feature of language use, but it is at the same time a potential source of unreliability.

The very unpredictability of such moods and emotions, however, means that there is a limit to the effectiveness of whatever measures we might take to deal with their effects. And if for some reason a learner does not feel like writing or talking, there is not a lot that we can do.

Predictable: There may be another set of affective factors which are predictable in their effects on the quality of communication, and which can therefore be built into a model of communicative performance. This is still an area of great ignorance and one worthy of much more research: we need to know what the predictable affective factors are, and what their sphere of influence is. It could be, for instance, that performance in spoken and written language is influenced by different sets of factors. But if we may from now on narrow our focus to performance in the spoken language, candidates for inclusion in the relevant set of predictable affective factors will include the age, status, personality-type (eg. 'out-going', 'reserved'), acquaintance-relationship, and gender of the participants. Let us now turn to three small studies which have aimed to shed some light on these questions, and to their implications.

4 Three Small Experimental Investigations

Investigation 1: Locke (1984) felt that the quality of spoken language elicited in an interview, or in any other face-to-face spoken interaction, might be crucially affected by features of the interlocutor - in the case of the interview, by features of the interviewer. Thus, if the interviewee was given interviewer 'a' he might do well, but if he was given interviewer 'b' he might do badly. Intuitively, her concern seemed reasonabie, and was backed up by a wealth of anecdotal evidence. Yet most testing concern with unreliability in interview assessment focuses on lack of consistency in the assessor; attempts to strengthen reliability in the assessment of speaking ability focus on assessor training and the use of adequate and appropriate rating scales. Whilst the latter are undeniably important, the more fundamental point that the quality of spoken language performance may vary predictably with features of the interlocutor tends to go unnoticed. Research in this area is practically non-existent, although the results would be of importance beyond language testing for our understanding of the nature of linguistic performance.

Locke chose to consider the effect of the gender of the interviewer on the interviewee. Four male postgraduate Iraqi and Saudi students at the University of Reading were each interviewed twice, once by a male and once by a female interviewer. The four interviewers w re all of comparable age. Two students were interviewed by a male interviewer first, and the other two by a female interviewer first; in this way it was hoped that any order effect could be discounted. Then, it was necessary for each interview to be similar enough to allow meaningful comparison of results, but not so similar that the second interview would be felt to be a simple repeat of the first, with a consequent practice effect. A 'same-but-different' format was therefore necessary. Each interview was given the same structure, and the general topic-area was also the same, but the specific content of the first and second interviews was different.

Each interview was video-recorded. Recordings were subsequently presented in a shuffled order, and assessed by one male and one female rater, each using two methods of assessment, one holistic (Carroll, 1980) and one analytic (Hawkey, 1982). In this way 16 comparisons of spoken language quality with male and female interviewers could be made.

Although the number of students was very small, the result was clear and provocative: there was an overwhelming tendency for students to be given higher ratings when interviewed by male interviewers. The tendency was evident in both scoring methods and there was a high level of agreement between the two raters.

Investigation 2: These results demanded both replication and deeper exploration. The writer therefore carried out a slightly larger investigation with thirteen postgraduate Algerian students at Reading (11 males and two females). This time, interviewers were cross-categorized not only by gender, but also by whether or not the student was acquainted with them and by a rough categorization of their personality as 'more outgoing' or 'more reserved'. Once again, the age of interviewers was comparable.

As in Locke's study, order of presentation was controlled for, with six students being given the female, and seven the male interviewer first. Cutting across the male-female category, as far as was possible (given the odd number involved) roughly half of the students were acquainted with the interviewer in the first interview, and unacquainted in the second, with the other half of the students having the reverse experience; and again roughly half of the students received an 'outgoing' interviewer first, followed by a 'reserved' interviewer, with the remainder having the reverse experience. The interviews were again designed to be 'same-but-different', were video-recorded, shuffled, and rated using two methods of assessment.

The tendency observed in Locke's study, for students to be rated more highly when interviewed by men, was once again overwhelmingly found. The tendency was equally clear in both scoring methods, and the degree of difference was fairly constant at about .5 of one of Carroll's bands. Interestingly, neither of the other potential factors considered - acquaintanceship and personality-type - could be seen to have any consistent effect.

What was not clear from Locke's study and could only be trivially investigated in this one was whether any gender effect was the result of interviewees' reactions to males versus females, or to own-gender versus opposite-gender interviewers. In this respect, it was particularly unfortunate that more female students could not be incorporated in the study: female students of the same cultural background as the males were not available. Nevertheless, while expressing all the caution necessary when considering the results of only two students, the results for the two female students were interesting. For one of the women, no difference was observable by either scoring method with the male and female interviewers. The other woman was rated more highly when interviewed by the man. Neither woman could be seen to go against the trend established in the men.

A very tentative conclusion to be drawn from these two limited studies would seem to be that, in the interview situation at least, young adult male Arab students may have a consistent tendency to produce a higher quality of performance in spoken English when being interviewed by a man than when being interviewed by a woman.

If these studies really have, in a preliminary way, succeeded in detecting a predictable affective factor in spoken language performance, a number of further questions will need to be researched to clarify just what that affective factor is. As has been suggested above, it is still not clear whether what has been observed concerns reaction to a male interviewer or to an own-gender interviewer. Further studies with female students would be needed in an attempt to answer this question.

Again, to what extent would this factor be restricted to Arab students? The emotive power of gender must surely pervade mankind, and thus such a gender-effect could be expected not only in any part of Europe but world-wide. On the other hand, Japanese colleagues say that they would not expect a gender-effect with Japanese students, but would not be surprised to find an age-effect, ie. we might expect students to achieve higher spoken-English ratings when interviewed by older interviewers, as such interviewers would be accorded greater respect. This interesting suggestion thus relates quality of performance in spoken language to the idea of degree of respect for the interviewer. A proposed gender-effect might thus be a manifestation of a more general 'respect' or 'status' effect. It might be that in many societies, but not all, men are accorded greater status than women, and that interviewees are moved to produce a higher quality of performance when confronted by high status in the interviewer. This suggests a need for a programme of research aimed at establishing and distinguishing between the effects of gender and status on quality of performance in spoken language.

Investigation 3: In an attempt to shed some light on this issue, a further small investigation was undertaken in Reading earlier this year. This is not yet complete, but preliminary indications are certainly of interest.

In this investigation, 16 postgraduate students were interviewed, coming from a variety of linguistic and cultural backgrounds. They included Arabs (Sudanese, Saudis, Yemenis and a Libyan), Japanese, Turks, and a Greek. Twelve students were male and four female.

As in the previous studies, each student was given two short 'same-but-different' interviews, one by a male interviewer, one by a female. Half of the students were interviewed by a male first, half by a female first, and all interviews were video-recorded.

The interviewers were roughly comparable in age, ranging from late twenties to early thirties. None of the interviewers was known to the students, and the personality of the interviewer was not controlled for. An attempt was made, however, to manipulate the status of each interviewer such that, in one interview the interviewer's status would be 'boosted' (high status), while in the next it would not be (neutral status). Each interviewer (I) interviewed four students, thus:

	1st interview	2nd interview
Student # 1	Male I # 1 High status	Female I # 1 Neutral status
Student # 2	Female I # 1 High status	Male I # 1 Neutral status
Student # 3	Female I # 1 Neutral status	Male I # 1 High status
Student # 4	Male I # 1 Neutral status	Female I # 1 High status

The status of an interviewer was manipulated in the following way: if status was being 'boosted' the interviewer was introduced to the student by family name, and with academic titles where relevant (eg. Dr Smith). A brief description of the interviewer's affiliation and most important responsibilities was given. Most interviewers in this condition wore some formal items of clothing (eg. jackets for both men and women, ties for men, etc.) and the person introducing the interviewers maintained physical distance between himself and them. An attempt was made by the introducer to indicate deference through tone of voice. If status was not being boosted - the 'neutral status' condition - interviewers were introduced in a very friendly way, by first name only, as friends of the investigator and sometimes as graduate students in the Department of Linguistic Science. Jackets, ties, etc. were not worn, and in each introduction physical contact was made between the introducer and the interviewer, in the form of a friendly pat on the arm. Interviewers were instructed to 'be themselves' in both status conditions, their status being suggested to the student purely through the mode of introduction together with minor dress differences.

Videos of these interviews are currently being rated on holistic and analytic scales, as before. On this occasion, however, the holistic scales used are those developed by Weir for the oral component of the Test of English for Educational Purposes (see Weir, 1988), and in order to facilitate comparisons, the videos have not been shuffled. Multiple rating is being undertaken, with an equal number of male and female raters. Thus far, only two sets of ratings have been obtained, one by a male rater and one by a female.

While it is as yet much too early to draw any solid conclusions, some tentative observations are possible.

Firstly, the two raters agree closely, on both rating scales.

Secondly, there is a slight tendency on both rating scales and with both raters for students to achieve higher ratings when being interviewed by males, but this is by no means as clear-cut as in the earlier investigations, and on the analytic scales there is considerable disagreement between the raters on which criteria, or for which students, this tendency manifests itself. Nevertheless, some tendency is there.

Finally - and this, perhaps is the most surprising finding - there is some slight tendency on the analytic scale, and a more marked tendency on the holistic scale, for students to achieve higher ratings with interviewers who were not marked for high status!

If this latter suggestion is borne out when the analysis is complete, and if it is reinforced when more substantial studies are undertaken, it will raise some perplexing questions of interpretation. One possibility might be that it is not rather specific factors such as 'gender' or 'age', and not even a rather more general factor such as 'status' which affect the quality of language production directly, but some much more general, very abstract factor such as 'psychological distance'. Thus the more 'distant' an interlocutor is perceived to be, the poorer the ratings that will be achieved. All kinds of secondary factors might contribute to this notion of 'distance', in varying strengths, but an interlocutor who is 'same gender', 'same age', 'known to speaker', 'same status', etc. might be expected to elicit higher-rated language than one who is 'other gender', 'older', 'unknown to speaker'. 'higher status', etc.

Whatever the primary and secondary factors which ultimately emerge, if the nature and degree of effect can be shown to be consistent in any way for a specifiable group of speakers, this will suggest that a gender or status or psychological distance feature will have a good claim to be incorporated in models of spoken language performance for those speakers, and that tests of this performance will need to take such predictable factors into account.

Let us now consider what such 'taking account' might involve, and finally relate the whole issue to our underlying concern with the complexity of tests.

5 Taking Account of A Predictable Affective Factor

It is certainly not widespread current practice to take account of gender, status of participants, or 'distance' between them, in tests of oral interaction. The selection of interviewer or other type of interlocutor is normally a matter of chance as far as such factors are concerned, and no attempt is made to adjust results in the light of them. Some post-hoc adjustment of ratings would of course be possible if the scale of an effect were known to be consistent. Thus a perfoımance rating with a male interviewer could be converted to an equivalent rating with a female interviewer, or vice versa. But we would now be touching on very sensitive matters. This should not surprise us, and is not a unique byproduct of the particular affective factor chosen by way of illustration; the reader is reminded that affective factors are matters of emotion and attitude, and it is not only the testee who is subject to their effects!

The question arises, then, of whether it is appropriate to adjust ratings in such cases. What would be the standard to which particular results would be adjusted? Many people feel that the test should give the learner the chance to show the best that he can do, with the implication that the test results should report the learner's best achievement. But what if that were to mean for many groups of male learners that spoken language achievement with a female interviewer would be converted to a predictive description of what they would have been able to achieve if they had been interviewed by a man? Or something between the two? For many, this would not be an acceptable solution.

A slightly different approach would be to recognize that humanity incorporates gender differences, status differences etc., and that the quality of linguistic performance is conditioned by such factors. Care should therefore be taken to allow all major relevant factors to have full and appropriate play in each component of a language test, and the description of performance which would be the output of the test would be understood to be based on an incorporation of such factors. Thus it might be appropriate for all interviewees to be multiply interviewed, by interviewers of varying degrees and types of 'distance'.

This type of solution would have the added attraction of being able to deal with the effects of affective factors in cases where it was predictable that the factors would have a marked effect, but not predictable how great or in what direction the effect would be. Thus a 'distance' effect might be great in some individuals, or in people from some cultural backgrounds, but slight in others; great 'distance' might depress the quality of performance in some learners, but raise it in others.

It might at first glance appear that such a 'full play' solution would also have the attraction of making it unnecessary to do the research to find out what the significant factors would be. Simply replicate as closely as possible those situations in which the learner would be likely to find himself, and the appropriate affective factors would come into play of themselves. However the practicality of test construction and administration will inevitably require some simplification of reality as it appears in the test, some selection of the features to include - including what is felt to be important, excluding what is felt to be irrelevant. Research into what the significant affective factors are, the scale of their effects, and their field of operation (what topic-areas, what cultural backgrounds, etc) will be necessary to inform the selection process.

6 Affective Factors and The Complexity of Tests

We have considered in this paper only one small area of affectiveness. There are certain to be others which affect language performance, perhaps of much greater magnitude in their impact. The spoken language only has been considered; it may be that some or all of the factors affecting the spoken language will be shown to have significant effects on performance in the written language, too, to the same or different degrees. Alternatively, there may be a quite different set of affective factors for the written language. And in both media, the term 'performance' may be understood to involve both reception and production. The potential for test complexity if all are to be reflected in test content, structure and administration is quite awesome. Even the 'full-play' proposal of the previous section, related to a 'status' or 'distance' effect alone, would double at a stroke the number of interviewers required in any situation. Nevertheless, a description of a learner's linguistic performance which ignored this dimension of complexity would be leaving out of account something important.

But yes, in the end, practicality will have to win the day. Where the number of people taking the test is relatively small, and where the implications of the results are not critical in some sense, it is unlikely that affective factors will be, or could be, seriously and systematically taken into account. But where the test is a large one, where the results can affect the course of lives or entail the expenditure of large sums of money, and where specifiable affective factors are known to have significant effects on linguistic performance, it would be dangerous to ignore them.

REFERENCES

Bachman, L. 1990. Fundamental Considerations in Language Testing. London: CUP.

Carroll, B J. 1980. Testing Communicative Performance: An Interim Study. Oxford: Pergamon.

Hawkey, R. 1982. Unpublished Ph D. Thesis, University of London.

Locke, C. 1984. Unpublished MA Project, University of Reading.

Oller, J W 1979. Language Tests at School. London: Longman.

Weir, C. 1988. Communicative Language Testing. University of Exeter Press.

AUTHENTICITY IN FOREIGN LANGUAGE TESTING

Peter Doyé

1. Validity of foreign language tests

Classical psychometric theory has taught us to evaluate the quality of educational tests by a number of basic criteria, such as validity, reliability, economy and utility. Although the characteristics of a good test can be classified in many different ways, test specialists are in general agreement that the criteria just named are the ones that any test producer or user should have in mind when making or applying a test.

They also agree that among the criteria mentioned above validity is the most important, for unless a test is valid it has no function. The validity of a test depends on the degree to which it measures what it is supposed to measure. A good test must serve the purpose that it is intended for, otherwise it is useless. However reliable the results may be, however objective the scoring may be, if the test does not measure what the test user wants to know it is irrelevant.

In our context most of the test users are foreign language teachers who want to know how well their students have learnt the foreign language. For this purpose they employ tests. My phrase "how well the students have learnt the foreign language" disguises the complexity of the task. In the past twenty or thirty years we have all learnt to accept communicative competence as the overall aim of foreign language instruction. Students are supposed to learn to understand and use the foreign language for purposes of communication. This general aim can, of course, be broken down into a number of competencies in listening, speaking, reading and writing.

In most countries the school curricula for foreign language instruction are formulated in terms of communicative competencies, and a logical consequence of this is that also testing is organized according to these competencies. This approach to testing has been called the "curricular approach". The foreign language curriculum is taken as the basis for the construction of foreign language tests. On the assumption that the actual teaching follows the content prescriptions laid down in the curriculum it seems plausible also to determine the content of the tests on the basis of the curriculum. This takes us back to the concept of validity. If the content of a test corresponds to the content prescribed by the curriculum it is said to possess "curricular validity" or "content validity".

2. Authenticity

However plausible the concept of content validity may be, in practice it presents a number of problems. One of these problems is the congruence of the test situation and the real life situation that the learner is supposed to master according to the curriculum. It is on this problem of congruence that I wish to concentrate in my talk. The problem has been described very aptly by Edward Cureton in his article on Validity in Lindquist's well-known book on Educational Measurement:

> If we want to find out how well a person can perform a task, we can put him to work at that task, and observe how well he does it and the quality and quantity of the product he turns out. Whenever a test performance is anything other than a representative performance of the actual task, we must inquire further concerning the degree to which the test operations as performed upon the test materials in the test situation agree with the actual operations as performed upon the actual materials in the situation normal to the task. One way to do this is to make detailed logical and psychological analyses of both the test and the task. From such analyses we may be able to show that many or most of the test operations and materials are identical with or very much like many or most of those of the task, and that the test situation is intrinsically similar to that of the task. On the basis of this demonstration it might be reasonable to conclude that the test is sufficiently relevant to the task for the purpose at issue. [1]

Let us try to apply the ideas expressed in this passage to a very common task that is to be found in any foreign language curriculum: Asking the way in an English speaking environment.

If we want to find out whether students are able to perform this speech act the safest way would be to take them to an English speaking town, place them in a situation where they actually have to ask the way and see whether they perform the task successfully and to which degree of perfection. We all know that this is hardly ever possible, except for language courses that are being held in an English speaking country. In the great majority of cases the teaching and learning of English takes place in a non-English environment. Therefore the second case mentioned by Cureton comes up when the tester tries to invent a realistic situation in which the learners have to perform operations congruent with the ones they would have to perform in situations normal to the task. Absolute congruence would exist when the tasks in the test situation and in the corresponding real-life situation would actually be identical. In this extreme case the test situation and the tasks in it are called authentic. An authentic test is therefore one that reproduces a real-life situation in order to examine the student's ability to cope with it.

There are authors who make authenticity one of the decisive characteristics of a good test. They derive it from the generally accepted criterion of validity and regard authenticity as the most important aspect of validity in foreign-language testing.

To quote just one author who takes this view: Brendan J Carroll:

> The issue of authenticity must always be an important aspect of any discussion on language testing. A full application of the principle of authenticity would mean that all the tasks undertaken should be real-life, interactive communicative operations and not the typical routine examination responses to the tester's 'stimuli', or part of a stimulus-response relationship; that the language of the test should be day-to-day discourse, not edited or doctored in the interests of simplification but presented with all its expected irregularities; that the contexts of the interchanges are realistic, with the ordinary interruptions, background noises and irrelevancies found in the airport or lecture-room; and that the rating of a performance, based on its effectiveness and adenuacy as a communicative response, will rely on non-verbal as well as verbal criteria. [2]

Brendan Carroll's whole book can be seen as one great attempt to ensure authenticity in language testing.

3. Limits to authenticity

It is at this point that I begin to have my doubts. However useful the postulation of authenticity as one criterion among others may be, it is certainly also useful to keep in mind that (a) a complete congruence of test situation and real-life situation is impossible and that (b) there are other demands that necessarily influence our search for optimal forms of testing and therefore relativize our attempt to construct authentic tests.

Re (a) Why is a complete congruence of test situation and real-life situation impossible? The answer is simple: because a language test is a social event that has - as one of its characteristics - the intention to examine the competence of language learners. In D Pickett's words: "By virtue of being a test, it is a special and formalised event distanced from real life and structured for a particular purpose. By definition it cannot be the real life it is probing."[3]

The very fact that the purpose of a test is to find out whether the learner is capable of performing a language task distinguishes it considerably from the corresponding performance of this task outside the test situation. Even if we succeed in manipulating the testees to accept the illocutionary point of a speech act they are supposed to perform, they will, in addition, always have in mind the other illocutionary point that is inherent to a test, namely to prove that they are capable of doing what is demanded of them.

An example of a test that examines the students' competence in asking for a piece of information: Even if by skillful arrangement we manage to lead the students to actually wanting this piece of information, they will always have another purpose of their verbal activity in mind which is: I will show you, teacher, that I am able to ask for information!

Re (b) The other obstacle on the way to perfect authenticity is an economic one. Through a test we want to get as much information about a person's communicative competence as possible. The greater the area of competence we cover by giving a particular test, the better. This requires a certain amount of abstraction from situational specifics. To use the example of Asking the Way: What we wish to know is how well the students can perform the speech act of Asking the Way in a variety of real-life situations - and the more the better - and not whether they can perform this act in the particular situation of a particular English city where they are looking for just one building in a specific street in a certain quarter of that city. However, we have to embed our task in a realistic setting that contains all these specifications in order to be plausible to the students. But this does not mean that we have to include all the incidentals that might be properties of such a real-life situation. On the contrary: the more incidentals we include, the more we move away from the general concept of Asking the Way as most of these incidentals might not be present in the majority of other situations where "asking the way" is demanded. Therefore we need not be sorry if we do not succeed in making a test situation absolutely authentic by providing all the peculiarities, background noises, hesitations, interruptions, social constraints by which a real-life communicative situation is characterized. We should endeavour to employ just the amount of realism that makes it understandable and plausible, but no more. The fact that we want to know how well the students master the essentials of our speech act requires abstraction from incidentals. Pickett gives the example of a simple arithmetic problem:

> If you are asked to find the area of a field 50 metres x 200 metres you do not have to get up and walk all over the field with a tape measure. You will not be concerned with whether it is bounded by a hedge or a fence, whether it is pasture or planted, whether it is sunny or wet or whether it is Monday or Thursday. These incidentals are irrelevant to the task of measurement, for which the basic information is ready to hand, and we know that the

> solution will not be affected by weather, time, cultivation, perimeter markings or any of the other factors which form part of our real-life perception of any particular field. The concept of area is an abstraction from all possible perceptions and is a constant.[4]

We have to concede that the decision about what are irrelevant incidentals is easier to make in the case of an arithmetic problem than in a communicative task, as communicative performance is always embedded in concrete situations with a number of linguistic as well as non-linguistic elements. But the arithmetic problem and the communicative task have one thing in common: Normally, ie., outside the artificial classroom setting, they occur in real-life situations that are characterized by a small number of essential features and a great number of incidentals which differ considerably from one situation to the next. And if we want to grasp the essential features of a task, we have to abstract from the incidentals. In this respect abstraction is the counterpoint to authenticity in testing.

What is needed is the right balance between authenticity and abstraction. We want a fair amount of authenticity but not so much as to obscure the essential properties of the speech act in question, which by virtue of being essentials obtain in all its manifestations. In this context, the findings of modern pragmatics can be of great help, I think. Its analyses of speech acts have demonstrated that every speech act has its own specific structure with certain characteristic features. It is on these characteristics that we have to concentrate if we wish to test the learners' competence in performing this particular act.

4. Examples

Let us take "Asking for Information" as an example. In his classical book "Speech Acts. An essay in the philosophy of language" John Searle has developed a systematic procedure for the description of speech acts, in which he presents the characteristic features of each act in terms of four kinds of conditions that are necessary and sufficient for the successful and non-defective performance of each act. The speech act of "asking for information" or simply "question" is one of the examples that Searle uses himself.

The essential characteristic of a question is that it counts as an attempt to elicit information from a hearer. The two preparatory conditions for the performance of a question are that the speaker does not know the answer and that it is not obvious that the hearer will provide the information without being asked. The propositional content of a question depends on what information the speaker needs, of course.

Now, we all know that teaching as well as testing the ability to ask questions is often practised in a way that disregards these conditions. A very common way is to present a number of sentences in which certain parts are underlined and to invite the students to ask for these parts.

Holburne Museum is situated in Pulteney Street.
It belongs to the University of Bath.
It is open daily from 11 am to 5 pm.
Mr. Green works in the Museum library.
He goes there every second morning.
He gets there by bus No. 32.
It takes him right to the main entrance.

This procedure is often used for the simple reason that it is easy to prepare, to administer and to score. But it very obviously violates the essential rules that govern the performance of a question. First of all, the speech act demanded cannot be regarded as an attempt to elicit information. Secondly, the testees do very well know the answer because it is given to them in the statements. It is even underlined, which normally means that the piece of information given is especially important - a fact that stresses the non-realistic character of the task.

And there is an additional negative feature: the procedure complicates the task for all those learners who find themselves incapable of imagining that they do not possess precisely the information that is given to them and to behave accordingly, ie. to pretend that they need it.

To conclude: The questions that the students have to ask in this test are no questions at all. The conditions under which they have to perform their speech acts are so basically different from those of real questions that the test cannot be regarded as a means to examine the students' competence in asking questions.

Let us look at the next example which could serve as an alternative to the previous one:

Holburne Museum is situated xx xxxxxxx xxxxxx.
It belongs xx xxx xxxxxxxxxx xx xxxx.
It is open daily xxxx xxxx xx xxxx.
Mr Green works xx xxx xxxxxx xxxxxxx.
He goes there xxxxx xxxxxx xxxxxxx.
He gets there xx xxx xxxxx.
It takes him right xx xxx xxxx xxxxxxxxx.

The difference between the two types of test is minimal on the surface, but decisive as regards the speech acts that are required to perform the task. By a very simple design, namely through replacing the underlined parts of the sentences by words that are illegibly written, the second type marks a considerable step forward in the direction of an authentic test: The questions that the learners have to ask are real questions in so far as the two main conditions of the speech act 'QUESTION' as elaborated by Searle are fulfilled.

First, they can be counted as attempts to elicit information and, second, the testees do not know the answers yet. What is still missing is an addressee to whom the questions might be addressed. Illegible statements are quite common, but one would hardly ever try to obtain the lacking information by a list of written questions. To make this test still more realistic, one could present the statements not in writing, but in spoken form with a muffled voice that fails to be clear precisely at those points where one wishes the students to ask their questions. In this case all the essential conditions of the speech act "QUESTION" would be fulfilled. But the test is still far from being authentic.

In a real life situation one would rarely find such a concentration of unintelligible utterances and therefore the necessity for a whole series of questions. Of course we can think of situations in which the necessity for quite a number of successive questions arises, such as in the situation of an interview or the situation of a game in which two partners need certain information from one another in order to complete a common task. - Two more examples are given.

5 The balance between authenticity and abstraction

But to come back to our central problem: How far do we want to go in our efforts to create authenticity?

In the middle part of my talk, I tried to explain why absolute authenticity, ie. complete congruence between the test situation and the so-called real life situation is neither possible nor desirable.

However much, for validity's sake, we might want to achieve authenticity in our tests, any attempt to reach it will necessarily arrive at a point, where it becomes clear that there are limits to authenticity for the simple reason that a language test - by its very purpose and structure - is a social event that is essentially different from any other social event in which language is used.

Very fortunately, we need not be afraid of protests from our students. They might be better motivated if we succeed in constructing tests that are highly authentic, for then they see the practical relevance of their tasks.

On the other hand most of them see as we do that a test can never become absolutely authentic and might find the vain attempts of their teachers to create fully authentic test situations fairly ridiculous. Therefore, and for the two main reasons I have presented we should give up our efforts to achieve the impossible and be satisfied with finding the right balance between authenticity and abstraction.

REFERENCES

1) *Cureton, Edward E: Validity In: Lindquist, E F (ed): Educational measurement. American Council on Education, Washington, D C. 1963, p. 622.*

2) *Carroll, Brendan J: Testing Communicative Performance Pergamon Press, Oxford, 1980. p. 11f.*

3 & 4) *Pickett, D: Never the Twain ...? COMLON Spring 1984. LCCIEB, Sidcup. P7.*

EVALUATING COMMUNICATIVE TESTS

Keith Morrow

1. THE CONTEXT: CCSE

In 1976, while working at the Centre for Applied Language Studies at the University of Reading, I was commissioned by the Royal Society of Arts to undertake a feasibility study into the development of a series of English language examinations based on the ideas about "communicative" language teaching which were then taking shape at Reading and elsewhere. 1. The outcome of this study was a series of examinations called the *Communicative Use of English as a Foreign Language* run by the RSA between 1980 and 1988, and subsequently run jointly by the RSA and the University of Cambridge Local Examinations Syndicate. I have recently completed a review of these examinations for Cambridge/RSA, and from November 1990 they are to be re-launched as *Certificates in Communicative Skills in English* (CCSE).

It will be clear that the origins of these examinations are to be found in a particular, though now widespread, view of what language, language teaching and language testing are about. In this paper I want to consider how this view has affected the design of the examinations, and then look at the question of the evaluation of the examinations in practice from the same perspective.

A number of characteristics of the new series of examinations relate directly and consciously to an underlying construct of what a "good" test ought to be.

1.1 Single skills

The examinations in fact consist of a suite of free-standing modules in *Writing, Reading, Listening and Oral interaction*. In each skill area, new tests are set at 4 levels for each series of the examination, and candidates are able to choose which modules at which level they wish to enter at any time. This structure reflects the experience of language teachers that the performance of students is not uniform across skill areas.

1.2 Tests of Performance

The tests are designed to be direct measures of performance. The justification from this again derives largely from an educational perspective on how tests should affect what is done in the classroom. In a "communicative" classroom, the focus of activities is (in simple terms) "doing something" through the language. We wanted to encourage this in washback terms, by designing tests which shared the same focus.

1.3 Task Based

More specifically, communication through language involves the participants in carrying out "tasks" in the production or comprehension of language in order to arrive at a shared understanding. This "negotiation of meaning" is most obviously a feature of face-to-face interaction, but it underlies all purposeful use of language whether in reading, writing, listening or speaking. The most striking implication of adopting this perspective on the design of a language test is the overriding importance of *authenticity* both of text (as input) and of task (in processing this input).

1.4 Criterion-referenced

The essential question which a communicative test must answer is whether or not (or how well) a candidate can use language to communicate meanings. But "communicating meanings" is a very elusive criterion indeed on which to base judgements. It varies both in terms of "communicating" (which is rarely a black and white, either/or matter) and in terms of "meanings" (which are very large and probably infinite in number). In other words, a communicative test which wishes to be criterion-referenced must define and delimit the criteria. This is a major undertaking, which for CCSE has led to statements for each of the four levels in each of the four skill areas of the *tasks* and *text types* which the candidates are expected to handle as well as (crucially) the *degree of skill* with which they will be expected to operate.

1.5 To reflect and encourage good classroom practice

Reference has already been made above to the educational effect of testing through the promotion of positive washback into the classroom. In the case of CCSE, this is a major concern underlying the design of the tests; indeed in many ways the tests themselves have drawn on "good" classroom practice in an attempt to disseminate this to other classrooms. This conscious feedback loop between teaching and testing, in terms not only of content but also of approach, is a vital mechanism for educational development.

2. EVALUATING TESTS OF THIS KIND

It will be clear from the preceding section that there is a conscious and deliberate rationale underlying the construction of the CCSE tests. However, a rationale can be wrong and some of the bases on which the tests are constructed may appear to be assertions ol what is believed, rather than outcomes of a process to determine what is right. Most significantly a "professional" language tester might perhaps wish to investigate the justification for the following design features.

2.1 Why tests of specific abilities not overall proficiency?

The idea of a single measure of overall competence or proficiency in a language is an attractive one for test designers and educational administrators. For administrators, the convenience of a single score or rating is obvious, and tests and exams as diverse as TOEFL, the Cambridge First Certificate, and the new Cambridge/British Council IELTS all manage to provide this. Why shouldn't CCSE? For test designers and researchers an "overall proficiency" model may also be attractive, not least for the scope it offers for writing papers reporting sophisticated investigations into the existence of this underlying "g" and the facters which may or may not contribute to it.

The CCSE scheme minimises the problem of weighting different factors by reporting performance in terms of each skill area individually. However, since within each skill area decisions have to be made about the contribution of individual tasks involving specific text types to the overall performance, there will still be scope for investigations of what makes up the underlying "l", "r", "s" and "w" factors in the listening, reading, speaking and writing tests.

The main justification for the apparent complexity of the structure of CCSE is not, in fact, to be found in the literature on language testing. Rather it is to be found in a deliberate attempt to provide on educational grounds an examination system which allows candidates to gain certification for what they can do without being penalised for what they cannot do. This is a stance which reflects a philosophical rather than an empirical starting point. Nonetheless, it is a common experience of teachers that many students have differential abilities in different language skill areas; it is important both practically and educationally that this is recognised.

2.2 Is there a conflict between authenticity and reliability?

The short answer to this question is probably "yes". In terms of test design, CCSE clearly needs to address itself to (at least) two areas of potential problem caused by the focus on the use of authentic tasks and texts. The first is the question of consistency of level of the tasks/texts used both within a particular test, and across tests in the same skill area in different series; the second is the question of consistency of judgement in the evaluation of performance, particularly in the *oral interaction* and *writing tests*.

In the implementation of the CCSE scheme, rigorous safeguards are in place to take account of these potential problems. A thorough moderating procedure is undertaken to scrutinise and trial tasks and papers to ensure consistency of content; and assessors for both the writing and oral tests are trained, briefed and monitored.

Yet still, it has to be said that the conflict remains. There are steps that could be undertaken to improve the reliability of the CCSE tests; but they would conflict directly with the authenticity criterion. Once again, it seems that in test design terms, the way that is chosen reflects a basic educational philosophy. From one standpoint, reliability is crucial; authenticity can be brought in to the extent that it is possible, but remains secondary. From another, the essential characteristic is authenticity; while recognising that total authenticity of task can never be achieved in a testing (or teaching) situation, every effort is made to focus on it. Reliability is not ignored and indeed every effort is made to ensure it receives due attention, but in the final analysis it is not the overriding factor in the design of the test.

It seems unlikely that in principle this conflict can be resolved. What is perhaps more important is to investigate how in practice the implementation of communicative exams like the CCSE proceeds.

2.3 How valid are the tests?

The third area of concern clearly focuses on validity. In the literature, language proficiency tests are investigated in terms of *construct, content and concurrent* validity, and *face* validity is also considered - sometimes as an afterthought, sometimes if nothing else can be found, but because it cannot be "measured", sometimes rather disparagingly. Some specific purpose tests are also evaluated in terms of *predictive* validity.

It will be clear from the preceding discussion that there is a "construct" to the CCSE scheme; the relationship between the construct and the content of any particular set of tests is open to empirical investigation - but unfortunately the construct itself is probably not. Similarly, in terms of content, the relationship between the specifications and the content of any particular set of papers can be investigated; but the more fundamental question of how far the specifications reflect the "real world" is not a matter of straightforward analysis. Concurrent validity is largely irrelevant because there are few other tests available against which CCSE can be measured. Face validity can be claimed - and claimed to be extremely important - but not proved. It may seem that CCSE should be open to investigation in terms of predictive validity since it is in essence making claims about the ability of candidates to carry out "real world" tasks. If they pass the test and can in fact carry out these tasks in the real world then the test may be said to have "predicted" this.

However, a moment's thought will show that this is in fact an impossible requirement. How would a researcher judge whether or not tasks had been "carried out" in the real world? The only way would be by evaluating performance on individual instances of the task - in which case all the problems of specification, reliability and generalisability which arise for CCSE would arise again. In a very real sense, the test would be validating itself against itself.

3. EPISTEMOLOGICAL TRADITIONS

In the preceding section we have seen how in three very basic respects, the design of a communicative test is based on factors which go beyond, or are not susceptible to, conventional language testing research: the overall design is founded on educational rather than testing requirements; reliability is secondary on construct grounds to authenticity; and the fundamental validity is not open to straightforward investigation.

This situation seems to raise some rather interesting questions about the kind of research which is appropriate to the investigation of language tests.

3.1 Language testing as pure science

Since the 1960's and the development of the "scientific" approach to testing, the predominant model of research has been one based on the precise measurement of data collected through empirical means. In other words, language testing has been seen as a branch of pure science, based essentially upon the twin concepts of quantification and analysis. In many ways this model has been extremely productive - at least of articles, books and PhD theses. But in terms of actually being able to make reliable, valid and comprehensible statements about what it is that tests measure, how this relates to language use, and how performance on tests relates to the real world, one might be tempted to echo the comments of a Nobel Prize winning economist about his own subject:

"In no field of empirical enquiry has so massive and sophisticated a statistical machinery been used with such indifferent results". 2

It seems to me that the reason for this sad state of affairs may well lie with the very notion of scientific rigour as it is generally understood, and the orthodoxy that this has imposed on our field.

"Under normal conditions the research scientist is not an innovator but a solver of puzzles, and the puzzles upon which he (sic) concentrates are just those which he believes can be both stated and solved within the existing scientific tradition" 3

At its best, this existing scientific tradition has encouraged researchers into what has become known as "McNamara's fallacy" of "making the measurable important instead of making the important measurable (or at least discernible)"4. At its worst, it might tempt us to paraphrase Oscar Wilde's epigram about the English upper classes and their fondness for fox-hunting. Wilde spoke of "The English country gentleman galloping after a fox - the unspeakable in full pursuit of the uneatable". Language testing researchers may not be unspeakable; but they may well be in pursuit of the unmeasurable. Elsewhere, 5 I have suggested that it may be time to propose a distinction between *language testing researchers* (who follow the existing orthodoxy) and *researchers into language testing* who would be prepared to adopt equally rigorous but rather different ways of looking at the area.

Support for alternative ways of conducting research into language testing seems to me to be available from two very different sources.

The first is the recent development (or at least the recent dissemination) of the ideas behind chaos theory. An exposition of this theory would be out of place here 6 (and in detail beyond my present understanding of it). But I find a set of ideas leads to the insight that conventional science finds it impossible to make a definitive statement about the length of a coastline (because of the problems of scale; the larger the scale, the longer the length because the more "detail" is included), or a firm prediction about the temperature of cup of "hot" coffee in a minute's time (because of the variability of convection), let alone what the weather is going to be like next week (because of the cumulative effect of a whole range of unpredictable and in themselves "trivial" events) an extremely powerful heuristic in thinking about language and language testing. Perhaps the key concept is "sensitive dependence upon initial conditions" - a way of saying that in looking at the world, everything depends on precisely where you start from. Nothing could be more appropriate for our field.

The second source of alternative ideas for the investigation of language testing comes from work carried out in the validation of experiential research. An extremely lucid account of this is given in Heron 1982. [7]. Experiential research is concerned above all to establish the *quality* or nature of the learning experience which participants have undergone. In this, it seems to me to relate very closely to important areas of concern for research into language learning and language testing.

Heron sets out a number of categories of validity which should be met by research. *Empirical* and *conceptual* validity can be related easily enough to categories which are familiar in our field; but his third category *ethical* validity perhaps opens up new areas. How far does existing research into language testing concern itself with the *ethics* of the test?

"Ethical validity has two aspects. Firstly is the research relevant to basic human concerns? Is it committed to values that make a difference to the quality of life for people now and in the future?...Secondly, do we behave morally while doing and applying the research...And do we deploy the results in ways that respect the rights and liberties of persons?" (Heron: 1982 p.1)

There are many questions raised here about the design and implementation of language tests, but perhaps the most obvious area of ethical involvement is the question of washback. It seems to me that a test which imposes (overtly or covertly) an impoverished or unrealistic classroom regime on students preparing for it is indeed "making a difference to the quality of life for people now and in the future" (though a difference of the wrong sort). This reinforces my view [8] that an important area of investigation in considering the validity of a test is an investigation of the classroom practice which it gives rise to.

A large part of Heron's paper is taken up with considering "Procedures for Distinguishing between the Veridical and the Illusory" in the context of experiential research. Again, this seems a rich field to harvest in considering the validity of language tests. This is not the place to consider all of these in detail, but one in particular is worthy of note. It is the principle of "authentic collaboration" between all the participants in a research project, breaking down "the traditional distinction between the role of the researcher and the role of the subject". As Underhill [9] points out, the use of label "subjects" is a widespread but de-humanising feature of most current work in language testing. In principle, the job of finding out what somebody is able to do in a language, or what the effectiveness of a particular test or test procedure is, might be greatly facilitated and would be much more "ethical" if the "subjects" were themselves genuinely involved in the process. Working out how to do this is of course another matter. But it does seem to point a new and very interesting direction for language testing and its associated research.

Underlying these two source., of new input to research on language testing, then, are ideas which move the centre of attention away from the conventional focus on the acquisition and analysis of "hard" data, to include a concern with those aspects of the nature of language itself which may not be susceptible to such procedures, and with the effects which tests and testing may have on the consumers. Testing and research which reflects this move will no longer be concerned simply to *measure*; rather it will establish a framework which will permit informed and consistent *judgements* to be made.

It would be unrealistic to claim that the CCSE scheme meets all the criteria (implicit and explicit) set out in this last section. But it is perhaps reasonable to claim that it illustrates the legitimacy of asking questions about tests other than those which language testing researchers conventionally ask. The next step is to find some answers.

NOTES

[1] *Techniques of Evaluation for a Notional Syllabus RSA (London) 1977. Now available from EFL Dept, UCLES, 1 Hills Rd, Cambridge, CB1 2EU UK.*

[2] *W Leontiev quoted in Chaos by J Gleick Cardinal 1988.*

[3] *Kuhn T S. The Essential Tension: Selected Studies in Scientific Tradition and Change, University of Chicago 1977 Quoted in Gleick 1988 (See note 2 above).*

[4] *Quoted in Assessing Students: How Shall We Know Them by D Rowntree. Open University 1977.*

[5] *See Ebvaluating Tests of Communicative Performance in Innovation in Language Testing ed. M. Portal Nelson/NFER 1987.*

[6] *For an excellent popular account, see Gleick 1988 (See note 2 above).*

[7] *John Heron Empirical Validity in Experiential Research Department Adult Education, University of Surrey, Guildford Surrey GU2 5XH*

[8] *See reference in Note 5 above.*

[9] *See N Underhill Testing Spoken Language Cambridge University Press 1988.*

MATERIALS-BASED TESTS: HOW WELL DO THEY WORK?

Michael Milanovic

INTRODUCTION

While all language tests tend to be materials-generating, their rationale and format is varied and they have differing effects on classroom practice. I would like to propose that language tests can be described as measurement-based, psycholinguistically-based and materials-based. Measurement-based tests tend to use a restricted item format, most commonly multiple-choice. They claim high reliability though are often criticized for lack of face and content validity. Psycholinguistically-based tests also tend to use a restricted range of item formats, such as cloze and dictation. It has been claimed that such tests tap an underlying language competence but they too have been criticized for lack of face and content validity. Materials-based tests arise out of trends in the development of language teaching materials. In recent years the most dominant generator of materials-based tests, in the British context at least has been the communicative language teaching movement. One important feature of materials-based tests is their use of a wide range of item formats which attempt to reflect teaching materials and currently, real-world language performance. Communicatively generated materials-based tests have tended to stress face and content validity but have placed less emphasis on reliability.

Materials-based test construction tends to be dynamic. New item formats are developed in line with developments in teaching methodology and materials. Measurement and psycholinguistically-based tests, on the other hand, tend to be static. The range of item formats does not change dramatically.

It is important to note that the distinctions made above are not clear cut. An item format may be materi^ls-based when it is first developed in that it represents current trends in teaching methodology or views of the nature of language competence. If it then becomes established, and continues to be used, despite changes in methodology or views of language, it is no longer materials-based. Ideally, tests should be materials-based, psycholinguistically-based and measurement-based concurrently. Only when this is the case, can we claim to have reliable and valid tests.

Hamp-Lyons (1989) distinguishes between two types of language testing research. The first is for the purposes of validating tests that will be operationally used. The second, which she calls metatesting, she defines as having its purpose in:

> *"... the investigation of how, why and when language is acquired or learned, not acquired or not learned, the ways and contexts in which, and the purposes for which, it is used and stored, and other such psycholinguistic questions".*

This type of language testing research has focused to a great extent on psycholinguistically and measurement-based test types and less on materials-based ones. In so doing, it has laid itself open to the criticism that too much attention has been paid to too restricted a range of item types. That not enough attention has been paid to understanding the interaction between background variables such as proficiency levels (Farhady, 1982) or the effects of the learning/teaching environment (Cziko, 1984) on test performance. The same might be said with regard to a systematic description of test content and the interaction between content and performance. Serious interest in this area is relatively recent (Bachman et al. 1988).

The aim of this article is to show that materials-based tests of English as a Second/Foreign language, reflecting both real-world and classroom language activities, can satisfy both measurement demands and provide interesting psycholinguistic insights. In other words, that there need not be an overpowering tension between the three perspectives outlined above. In practical terms, the tests and procedures used as examples here are most directly relevant in the context of a language teaching institute.

Test constructors, educators and test consumers need to be satisfied that tests are measuring what they are intended to measure consistently and fairly. Tests must be reliable because people's lives may depends on the results. For a variety of reasons it appears to be the case that many test construction agencies have been too willing to believe that satisfactory measurement criteria can only be achieved in a limited number of ways. In language testing, although this is also true in many other subject areas, this belief has led to the development and very wide use of indirect methods of testing ability. The most common such method is the multiple-choice item. It satisfies the conditions of objectivity of marking, economy of scoring and readily lends itself to statistical validation procedures. However, it does not have a very good effect on classroom practice, nor does it reflect the way language is used in real-world contexts.

A tension exists in the language teaching/testing world between the need for accountability in the educational process and the need to be accountable for the effects of testing on the educational process. In other words, while we must be able to trust the testing instruments that we use, it must be accepted that tests have a major influence on what goes on in the classroom. Both teachers and students generally believe, and rightly so to a great extent, that one of the best ways to prepare for a test is to practice the items in the test. It is a well established fact that the multiple-choice test format does not inspire innovative methodology, that it has had a largely negative effect on classrooms all over the world. Unhappily, it is still widely considered the best testing has to offer because it satisfies the need for measurement accountability and is economical to administer and mark.

In test validation research the problem of relating testing materials to useful and beneficial teaching materials has led to investigations of different test formats. Swain (1985) describes a Canadian project in which students actually participate in the creation of test items based on their own perceived needs. Swain formulates four principles that should guide the test constructor. These are:

i start from somewhere;
ii concentrate on content;
iii bias for best;
iv work for washback.

The first principle, start from somewhere, suggests that the test constructor needs to base test development on a model of language ability. The second principle, concentrate on content, suggests that test content should motivate, be substantive and partially new, that it should be integrated, and that it should be interactive. The third principle, bias for best, demands that tests should aim to get the best out of students, rather than the worst. Swain feels that it is important to try and make the testing experience less threatening and potentially harmful. The fourth principle, work for washback, requires that test writers should not forget that test content has a major effect on classroom, practice and that they should work towards making that effect as positive as possible. Clearly, these four principles cannot be satisfied by using only indirect measures such as multiple-choice items. We have to turn towards other item types.

There have been attempts originating from testing agencies to make language tests more relevant and meaningful. The Royal Society of Arts (RSA) in the United Kingdom developed a series of examinations in the Communicative use of English in the late seventies based on criteria proposed by Morrow (1979). The tasks appearing in these examinations attempted to reflect authentic communication activities and current trends in language teaching methodology. Great emphasis was placed on the involvement of language teachers in test construction and marking, and the backwash effect of this process, as well as the examinations themselves, on the teaching of English. It must be said that these are powerful features of the approach taken by examining boards in Britain. Examinations are not perceived as the property of boards alone. Ownership is distributed between the boards, methodologists and teachers, all of whom accept responsibility for the effect that the examinations have on the consumer - the students taking examinations - and the educational process. Many examining boards in the United Kingdom try to reflect language in use in many of the item types they use. This has been done in response to pressure from teachers demanding an approach that reflects more closely recent trends in methodology. The trend towards more realistic test items has not always been backed up by the equally important need to validate such tests. The combination of innovation and appropriate validation procedures is a challenge yet to be fully faced.

Even so, the examples cited above show that parts of the testing world are trying to move towards tests that look more valid and try to reflect both real life language activities and recent trends in language teaching methodology and materials more closely.

A major strength of the materials-based approach is that it actively works for positive washback effect. This helps to indicate to students, as well as teachers, that the main purpose of language instruction is to prepare students for the world outside the classroom. This should give the materials-based approach significant motivational value. However, as Wesche (1987) points out with regard to performance-based test construction (and the same is surely true with regard to materials-based tests):

> *"Performance-based test construction requires considerable advance or 'front end' work: careful specification of objectives, identification and sampling of appropriate discourse types, content and tasks, and consideration of scoring criteria and procedures."*

When preparing materials-based tests, achieving reliability may appear to be difficult due in part to the untried nature of many of the item types and in part to the fact that achieving reliable measurement is always a problem. However, both reliability and validity have to be established. Extensive investigation, moderation and pretesting procedures have to be employed to achieve both reliability and validity at the expense of neither.

While several attempts have been made to produce face, and to some extent content valid language tests, a disturbing lack of attention has been paid to making such tests reliable, or establishing their construct validity. In the following I will describe a project that attempted to produce a test battery that was based, to some extent at least, on the real world needs of the test takers. It took place in the British Council language teaching institute in Hong Kong.

The British Council language institute in Hong Kong is the largest of its kind in the world. There are between 9,000 and 12,000 students registered in any one term. In the region of 80% of the students are registered in what are loosely called General English courses. In fact this term is misleading. Through a fairly standard ESP type of investigation into the language needs of the students, it was possible to show that two main categories of student were attending courses. These were low to middle grade office workers, and skilled manual workers. This meant that the courses could be designed with these two main categories in mind. A much smaller third category was also identified, though this overlapped heavily with the first two. This category was students learning English for varied reasons. A set of real-world English language performance language performance descriptions were generated. These formed the basis for test specifications and the generation of teaching materials.

TEST CONTENT

An achievement or progress test should reflect course content. This is not to say that each item in the course needs to be tested. Unfortunately, in the minds of many teachers and students a test needs to cover all aspects of a course to be valid or fair. If the test is a discrete-point grammar test, testing a discrete-point grammar course then this may be possible if not desirable (Carroll, 1961). In almost any other context it is simply not possible to test all that has been taught in the time available for testing. In deciding test content the following points need to be considered:

i A representative sample of areas covered in the course need to appear in the test. (the term 'representative' is not defined accurately. Its meaning will vary from context to context, and test to test);

ii Enough variety needs to be present to satisfy teachers and students that no one is being discriminated against or favoured in any way;

iii The item types that appear in a test must be familiar to both teachers and students.

iv The test content must not appear to be trivial.

v There must not be an undue emphasis in the test areas of minor importance.

vi The use of item formats suited primarily to testing purposes eg. discrete-point multiple-choice, should be avoided as far as possible if they conflict with sound teaching principles (whatever these may be).

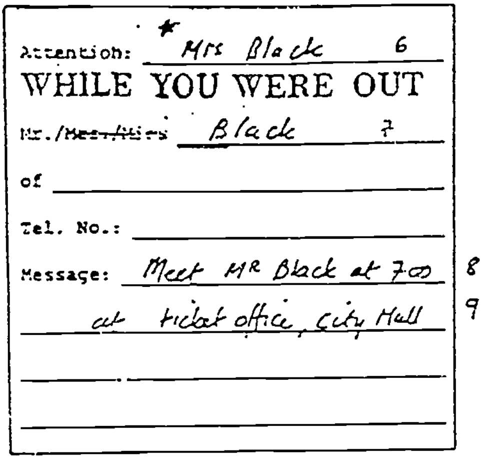
All too often operationally used tests do not resemble teaching materials in style and format. If, teaching a language aims to prepare learners for real-world use of that language then it is reasonable to assume that certain tasks encountered in the classroom will, to some extent, reflect reality. Other tasks may be of a purely pedagogical nature. There must, for students and teachers, be either a pedagogical or real-world familiarity with items in a test - preferably both.

Items to be included in tests should be selected on the basis of their relevance and familiarity and the extent to which they are, when incorporated into a test, reflective of the course students followed and the ways in which they put language to use.

TASK-BASED VS DISCRETE-POINT ITEMS

The argument above raises the question of whether test items should be task-based or discrete-point. As teaching becomes more whole-task-based it is inevitable that test items must follow. However, this causes two sets of problems from a testing point of view. Firstly, how is the tester to sample effectively from all the task-based activities and to what extent are the results obtained generalizable? These problems have been discussed at length over the years but no satisfactory solution has been reached.

Secondly, in real life, a task is generally either successfully completed or not. In class, the teacher can focus on any aspect of the task in order to improve student performance. In the testing context, however, the task may provide only one mark if treated as a unity, as long as an overall criterion for success can be defined and whether this is possible is a moot point. Such a task may take several minutes or longer to complete. If the test in which it resides is to be used for ranking or grading it can be extremely uneconomical to treat a task as a single unit. An example of a task based item would be the telephone message form illustrated below.



Attention: * Mrs Black 6

WHILE YOU WERE OUT

Mr./~~Mrs./Miss~~ Black 7

of

Tel. No.:

Message: Meet MR Black at 7:00 8

at ticket office, City Hall 9

6.	1	0	0
7.	1	0	0
8.	1	0	0
9.	1	0	0

Clearly, for the task to have been successfully completed all the relevant information needs to be present. Unfortunately this is rarely the case - mistakes are made, information is missing. It would be difficult to score such an item dichotomously and achieve a reasonable distribution of scores or provide enough information for effective test validation.

A compromise solution that satisfies the criterion of authentic/realistic appearance, allows the tester to allocate an appropriate number of points to the task to make it economical from a scoring point of view, and provides relevant data for validation, is to break a task down into discrete points for marking purposes. It is important the student does not perceive such a task as a group of individual items but rather as a whole task.

CONSULTATION IN TEST CONSTRUCTION

The views of both students and teachers are important in test construction. It is difficult to involve students in test construction, but it is of great importance that their views are sought after pre-testing or test administration in order that objectionable items can at least be considered again. It is often enough for teachers to ask for informal feedback at the end of a test. Some recent research has also focused on introspection by students.

Equally important as the views of the students is that of the teachers. At best the concept of testing in English language teaching is unpopular and badly understood. For any approach to testing to succeed, therefore, three factors are of vital importance:

i Teachers must gain some familiarity with the principles and practice of language testing. This is perhaps best achieved through some form of basic training course;

ii Teachers must be involved in the process of test design, item format selection, and the writing of test items;

iii Teachers must be familiar with the life cycle of a test and aware of the ıct that good test construction cannot be haphazard.

It is unfortunately very difficult to achieve any of the three aims in a short period of time with an entire teaching body of any size. In the case of the British Council institute in Hong Kong, there were more than one hundred teachers employed at any one time and so, training and involvement had to take place by degree. However, it was anticipated that the credibility of the tests and the process of consultation would be better accepted when those who were actually involved in working on the tests mixed with teachers who were not involved. The more teachers could be made to feel a personal commitment to the tests, the more people there were who would be available to explain and defend them as necessary. The image of the test constructor in the ivory tower having no contact with the teaching body had to be dispelled as fully as possible. Thus it was that there were generally between four and six teachers involved in test construction in any one term.

A MATERIALS-BASED TEST

One of the tests in the battery developed in Hong Kong will now be described in order to illustrate some of the points made earlier. The A3 Progress test, like all the others, is divided into four basic parts. A3 level students have a fairly low standard of English therefore the test tasks they have to perform are of a rather basic kind. Every attempt was made, however, to keep these tasks realistic and relevant.

The Listening Test, a copy of which appears in appendix 1, comprises three item types. The first simulates a typical telephone situation that the students are likely to encounter, the second a face to face exchange at a hotel reception desk, and the third a face to face exchange between a travel agency clerk and a tourist booking a day, tour. The skills tested are listed below:

Taking telephone messages

This involves:

- writing down spelling of names;
- writing down telephone numbers;
- writing down short messages (instructions, places, times).

Writing down information about a customer

This involves:

- writing down spelling of last time;
- writing down first name when not spelt;

- writing down 'Tokyo' (not spelt);
- writing down spelling of address;
- writing down name of local airline (not spelt).

Writing down information for customers at a travel desk

This involves:

- writing down spelling of name;
- writing down room number;
- writing down number of people going on trip;
- writing down times of day;
- writing down price.

In the real world, skills frequently tend to integrate. This feature of language use was accepted as fundamental to item design. However, it should be noted that reading and writing are kept to a minimum in the Listening test. It was felt that it would be unfair to include a significant element of either of these two skills, since the students' competence in both was likely to affect performance in listening. Enough reading and writing was retained to ensure the reality of the tasks while not hindering students in their completion of these tasks. The tape recordings were made in studio conditions and various sound effects incorporated to make them more realistic.

The Grammar Test caused some concern. It was decided that the tests should include a section on grammar, or perhaps more appropriately, accuracy. The communicative approach has been much criticized by teachers and students for its perceived lack of concern for the formal features of language. In the Hong Kong context, it was very important to the students that there should be something called grammar in the tests. From the theoretical point of view, it was also felt that emphasis should be placed on more formal features of language. How they should be tested was the difficult question. If standard discrete-point multiple-choice items were used, the washback effect on the classroom would have been negative in the sense that the multiple-choice approach to grammar teaching was not a feature of the teaching method in the British Council. It was also thought better to use an item type which was text-based as opposed to sentence-based. To this end a variation on the cloze procedure was developed for use in the lower level progress tests. It was given the name 'banked cloze' because, above each text, there was a bank of words, normally two or three more than there were spaces in the text. Students chose a word from the bank to match one of the spaces. Each text was based on some authentic text-type relevant to and within the experience of the students. These were:

An article from Student News.
A newspaper article.
A description of an office layout.
A letter to a friend.

It should be pointed out that the same format was not used at higher levels. A method of rational deletion (Alderson, 1983) was used instead. It was accepted that there were many potential hazards in the use of the cloze. However, it satisfied the washback requirements better than any other item-type available at the time.

The Appropriacy Test was based on the common teaching technique, the half and half dialogue. Situations relevant to and within the experience of the students were selected. One person's part of the dialogue was left blank and it was up to the student to complete it as best he could. Clearly, writing down what would be said in a conversational context suffers from the point of view that it is not very realistic. However, it was a teaching device commonly used in the institute, and thus familiar to the students. Furthermore, it focused attention on the sociolinguistic aspects of language and allowed for a degree of controlled creativity on the part of the student. The marking was carried out on two levels. If the response was inappropriate it received no marks, regardless of accuracy.

If it was appropriate, then the marks were scaled according to accuracy. Only a response that was both appropriate and wholly accurate could receive full marks.

The types of functional responses that the students were expected to make are listed below:

- giving directions;
- asking about well being;
- offering a drink;
- asking for preference;
- asking about type of work/job;
- asking about starting time;
- asking about finishing time;
- giving information about own job;
- giving information about week-end activities.

Reading and Writing were the final two skills areas in this test. An attempt was made here to integrate the activity as much as possible, and to base the task on realistic texts. Students were asked to fill in a visa application form using a letter and passport as sources of information. The passport was authentic reading material, while the letter was especially written for the test. The form was a slightly modified version of a real visa application form. The introduction of authentic materials into the test as opposed to contrived teaching materials, and a focus on a situation that any of the students may need to deal with was an important statement. The test was attempting to do something that, at the time, most of the teachers were not, that is, using authentic materials with low proficiency students. The teachers soon saw that the nature of the task was as important as the material. They were also able to see that students almost enjoyed this sort of activity, and immediately understood its relevance to their day-to-day lives. Informal feedback from teachers, after the introduction of the test, indicated that it had encouraged a greater focus on the use of authentic materials and realistic tasks in the classroom. It seemed that positive washback was being achieved.

THE TEST CONSTRUCTION PROCESS

Little guidance has appeared on how to actually develop a communicative test battery or integrate it into the workings of a school environment. Carroll (1978; 1980) gives the matter of test development some coverage but he does not consider, in any depth, the consequences or role of testing in an educational context. With regard to involving teachers and integrating testing into the school environment, there is also very little guidance available. Alderson and Walters (1983) discuss the question of training teachers in testing techniques on a postgraduate course. The process of training and sensitization in-service is not considered.

Inextricably linked to the process of test development, as described here, is the need to actively involve and train teachers in the institute in test design and implementation. The tests developed in Hong Kong underwent very similar treatment before they were finally implemented. It was through involving teachers in the stages of this treatment, that some degree of training and sensitization was achieved. Listed below are the six stages of test preparation. I believe they are appropriate to many situations where teaching and testing interact.

Stage 1

Test construction needs to be coordinated. At the beginning of a test construction cycle, the testing coordinator needs to meet with a group of test item writers, normally teachers, specializing in writing items for a given test. In this case 'specializing' means teachers who have worked with students at a given level and are preferably teaching them. The purpose of a preliminary meeting is to discuss any ideas that the teachers may have, to take into account any feedback regarding the tests already operating and decide on a topic area that each teacher could focus on in order to prepare items for the next meeting. Teachers need to be briefed on some of the difficulties they are likely to encounter in test item writing, and how they might cope with such difficulties.

Stage 2

The teachers write first draft items in light of Stage 1 discussions, their experience of the materials and students, the course outlines and performance objectives.

Stage 3

A series of meeting is held when the items prepared by individual teachers are subjected to group moderation. The items are discussed in terms of their relevance, testing points, importance, and suitability for the students in question. It is important that any idiosyncrasies are removed at this stage.

Group moderation is a vital phase in the preparation of items for several reasons. Firstly, in test construction, where great precision and clarity are required, several people working on an item inevitably produce better results than just one person working alone. Secondly, a group product is generally better balanced and more widely applicable if worked on by teachers all actively engaged in teaching a course. Thirdly, the teachers in the test construction team are well prepared for many of the questions that might later arise from the use of a particular item and are able to justify its inclusion in a test.

Teachers are often found to rush moderation at first because they may be worried about offending their colleagues or unable to focus precisely enough on the likely problems or difficulties an item may pose, such as markability, reasonable restriction of possible answers and so forth. It is important to insist on thorough moderation at this stage since without it the product will probably be of inferior quality and may need complete re-writing and pretesting before it is of any use.

Stage 4

Completed items are then informally trialled with participating teachers' classes in order to uncover any glaring difficulties that the moderation team had not been able to predict. This helps to greatly increase the sensitivity of teachers engaged in item writing. It is all too commonly believed by teachers and administrators alike that test construction can be accomplished quickly and that the product will still be quite acceptable. Unfortunately, due to a number of factors such as the unpredictability of the students, the shortsightedness of the test writer, the lack of clarity in instructions, this is rarely the case. Initial moderation helps to make teachers aware of some of the difficulties; trialling informally with their own classes is an invaluable addition to this sensitization process. Moreover, teachers have the opportunity of observing the reactions of students to the items and the way in which they attempt to do them. Both of these factors are very important in the construction of task-based tests that attempt to have a positive washback effect on the classroom.

Enough time needs to be allocated to Stages 1-4. In the context of a teaching institution, given the range of demands on everyone's time, at least three or four months is required for the successful completion of these stages.

Stages 5

After initial trialling, the moderation team meets again, and in light of the experience gained so far prepares a pretest version of a test or part of a test.

The pre-test is then administered to a representative sample of the population and the results analyzed. It is generally necessary to pre-test up to twice as many items as will eventually be required to achieve the appropriate quality.

Stages 6

The moderation team meets to discuss the results of the pretest and decide on the final form of the test items.

Any test item generally takes at least six months from inception to completion in the context under discussion here. Teachers should be involved in the process from start to finish. Those teachers involved realize that the process of test construction, while lengthy and time consuming, must be carried out with the greatest of care because the test results have a very real influence on the students in question. They are able to bear witness to the fact that no test can be produced without due care and attention. To begin with, most of them believe the approach to be unnecessarily long drawn out and tedious, but as they work on items and become fully aware of the fallibility of tests and test constructors, their attitudes change.

Do these tests meet measurement criteria:

I made the claim earlier that materials-based tests need to function at least as well as measurement-based tests, from a statistical point of view. Even if the same degree of economy of marking cannot be achieved, this is out weighed, in an institutional context, by the considerable educational benefits.

Some basic test statistics for five progress tests from the battery in question are presented below. Each test was analyzed in two ways. Firstly, it was treated as a unity, in the sense that none of the sections were analyzed separately. This means that the mean, standard deviation, reliability and standard error of measurement were established for the whole test. Then each section was treated as a separate test. This meant that there were four separate analyses of Listening, Grammar, Appropriacy, and Reading and Writing.

Table 1.

	WT	LIS	GRM	APP	RD/WT
A3 Test					
$\bar{X}$	63%	55%	60%	81%	69%
SD	19%	24%	22%	24%	28%
KR20	0.95	0.92	0.88	0.84	0.92
NQ	89	28	29	10	22
NS	264	264	264	264	264
B1 Test					
$\bar{X}$	54%	42%	52%	77%	53%
SD	16%	20%	21%	18%	28%
KR20	0.93	0.87	0.83	0.78	0.89
NQ	96	33	24	19	20
NS	305	305	305	305	305
B2 Test					
$\bar{X}$	58%	42%	57%	74%	65%
SD	14%	18%	18%	15%	19%
KR20	0.91	0.82	0.80	0.68	0.85
NQ	99	29	24	20	26
NS	259	259	259	259	259
C1 Test					
$\bar{X}$	57%	55%	46%	80%	64%
SD	16%	20%	19%	23%	24%
KR20	0.94	0.88	0.86	0.84	0.91
NQ	112	34	35	12	31
NS	250	250	250	250	250
C2 Test					
$\bar{X}$	58%	57%	49%	79%	62%
SD	18%	20%	21%	22%	27%
KR20	0.95	0.86	0.87	0.74	0.91
NQ	98	31	31	09	25
NS	242	242	242	242	242

KEY

WT	=	Whole Test
LIS	=	Listening
GRM	=	Grammar
APP	=	Appropriacy
RD/WT	=	Reading and writing
X	=	mean score;
SD	=	standard deviation
KR20	=	Kuder-Richardson 20 reliability quotient;
NQ	=	number of items in the test or subtest;
NS	=	number of students in the sample

Table 1 illustrates basic overall test and subtest statistical characteristics.

It is clear from these figures that the tests are very reliable. The reasons for this are as follows:

i. much time and effort was put into planning and moderation;

ii. test content was relevant and well defined;

iii. teachers were involved in the process of test writing from the earliest stages;

iv. the tests were all pretested and revised in light of pretest performance.

Do these tests meet psycholinguistic criteria?

Meeting psycholinguistic demands is a complex issue at several levels. In this context, the most straightforward of these is to attempt to show that the subtests are indeed measuring different aspects of underlying language performance. In order to do this it is necessary to demonstrate that tasks in a subtest relate to each other more closely than they do to tasks in other subtests. The most widely used methodology to investigate this type of issue is factor analysis. Simply put, factor analysis is a correlational technique which attempts to reduce the number of observed variables to a smaller number of underlying variables. It does this by grouping the observed variables on the basis of how closely related they are to each other. It is then up to the researcher to interpret the findings.

In the case of the tests in the battery described here this was done by computing students' scores on subtest tasks and then treating these tasks as mini-tests in their own right. If the tasks grouped together according to the skills they were said to be testing, then this would provide evidence that performance could be accounted for by different underlying skills. A factor analysis for the A3 test is illustrated in Table 2.

Table 2.

Subtest	Factor 1	Factor 2	Factor 3	Factor 4
Listening 4	.74541			
Listening 1	.70287			
Listening 5	.64940			
Listening 2	.64851			
Listening 6	.63182			
Listening 3	.62097			
Grammar 1		.75096		
Grammar 4		.69953		
Grammar 2		.63289		
Grammar 3		.51338		
Approp 1				
Rd/Wrt 4			.86169	
Rd/Wrt 2			.65637	
Rd/Wrt 5			.59547	
Rd/Wrt 3		.41049	.52075	
Rd/Wrt 1			.44136	
Approp 2				.75125
Approp 3		.41395		.54720

Interestingly, at this fairly low level of proficiency, it is clear that subtest tasks are more closely related to tasks testing the same skill than they are to tasks testing other skills. There is a very clear differentiation between the skills. Most experienced teachers would not find this discovery startling. In the lower and intermediate stages of language acquisition learners clearly develop skills differentially. In other words, a learner may be good at listening and bad at reading. Analyses of tests are different levels of proficiency is reported more fully in Milanovic (1988). The findings of this research indicated that, as learners' language proficiency increased, the skills tended to merge more with each other. A similar finding has been reported by de Jong (1990) using Rasch analysis as opposed to factor analysis. Such evidence casts doubt on the findings of language testing research that does not take the proficiency level of learners into account.

CONCLUSION

The results and procedures described here show that materials-based tests can work. In an educational context, where possible, such tests should be used in preference to approaches further removed from the classroom or real-world context. They are educationally far more desirable than more traditional tests and lose nothing in terms of reliability, if well prepared. In addition, it is time that more innovative tests formed the basis for research in language testing. They would be a more relevant starting point than tests that reflect thinking thirty years ago.

Canale (1985) amongst others, has pointed out that there is often a mismatch between teaching/learning materials and those that appear in proficiency -oriented achievement tests. He attributes the mismatch to what he calls the 'image problem', which he breaks down into several categories. First he focuses on the role of the learner in testing and describes him as typically:

> *"an obedient examinee, a disinterested consumer, a powerless patient or even an unwilling victim".*

Canale also focuses on the type of situation that current achievement testing often represents:

> *"... it is frequently a crude, contrived, confusing threatening, and above all intrusive event that replaces what many learners (and teachers) find to be more rewarding and constructive opportunities for learning and use".*

The problems that Canale outlines, which are also of concern to Swain (1985), are major difficulties in the acceptability of testing as an important and useful part of the educational process. Several strategies can be adopted to overcome these problems.

Firstly, testing programmes should be integrated into the life of the institution in which they occur. Testing specialists need to be involved in all stages of curriculum design and not seen as additional extras to the process.

Secondly, the materials used in tests should always reflect the types of activities that go on in the classroom and/or the lives of the students taking the test. In this way both teachers and students will have the better chance of seeing the relevance of tests.

Thirdly, teachers' sometimes inadequate understanding of testing purposes, procedures and principles are often a major barrier in the successful integration of testing into the curriculum in order to overcome this problem, teachers need to be actively encouraged to get involved in test writing projects, and there needs to be a heavy emphasis on their training. Such a strategy not only improves the quality of tests, in terms of reliability and validity as illustrated earlier, but also means that more teachers will become familiar with testing as a discipline that is integrated into the education process and not apart from it.

BIBLIOGRAPHY

Alderson, J C., 1983. The cloze procedure and proficiency in English as a foreign language, In Oller, J W (ed), 1983.

Alderson, J C and A Waters, 1983. A course in testing and evaluation for ESP teachers, or 'How bad were my tests?' In A Waters (ed), 1983.

Bachman L F, et al. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries, Language Testing V5 No 2 pp 128-159.

Canale, M. 1985. Proficiency oriented achievement testing. Paper presented at the Master Lecture series of the American Council on the teaching of Foreign Languages, at the Defence and Language Institute.

Carroll J B. 1961. Fundamental consideration in Testing for English proficiency in foreign students, in Testing the English Proficiency of foreign students, CAL, Washington DC, pp 31-40.

Carroll, B J. 1978. Guidelines for the Development of Communicative Tests, Royal Society of Arts, London.

Carroll, B J. 1980. Testing Communicative Performance, Pergamon Press, Oxford.

Cziko, G. 1984. Some problems with Empirically-based models of communication competance Applied Linguistics, Vol 5 No 1.

Farady, H. 1982. Measures of Language Testing proficiency from the learners' perspective, TESOL Cuarterly, Vol 16 No 1.

Hamp-Lyons, Liz (1989a). Applying the partial credit method of Rasch analysis: Language Testing and Accountability, Language Testing, Vol 6:1, 109-118.

Milanovic, M. 1988. The Construction and Validation of a Performance-based Battery of English Language Progress Tests Unpublished PhD Thesis, University of London.

Morrow, K. 1979. Communicative language testing: revolution or evolution, In Brumfit, C J and K Johnson (eds), 1979.

Swain, M. 1985. Large-scale Communicative language testing: A case study in Lee, Y.P. et al (eds), New Direction on Language Testing, Pergamon, Oxford.

Wesche, M B. 1987. Second language performance testing: the Ontario test of ESL as an example, Language Testing, Vol. 4, No. 1, 28-47.

DEFINING LANGUAGE ABILITY: THE CRITERIA FOR CRITERIA

Geoff Brindley

INTRODUCTION

In recent years, there has been a move towards the wider use of criterion-referenced (CR) methods of assessing second language ability which allow learners' language performance to be described and judged in relation to defined behavioural criteria. This is in line with the concern among language testers to provide meaningful information about what testees are able to do with the language rather than merely providing test scores. However, while criterion-referencing has enabled language testers to be more explicit about what is being assessed, there are numerous problems associated with the development, interpretation and use of CR methods of assessment, to such an extent that the feasibility of true criterion-referencing has been questioned by some writers (eg. Skehan 1984, 1989).

This paper aims to illustrate and discuss the nature of these problems in the context of both standardized proficiency testing and classroom assessment. First, different interpretations of "criterion-referencing" will be examined. Following this, a range of approaches to defining criteria and performance levels in second language assessment will be outlined and some of the issues which have arisen in defining and applying these criteria will be discussed, including the difficulties of defining the nature of "proficiency" and the failure of expert judges to agree on criteria. Finally, research directions will be indicated that might lead to language assessment criteria which incorporate multiple perspectives on learners' communicative needs and which derive from empirical data on second language acquisition and use.

CRITERION-REFERENCING

The term "criterion-referenced" has been interpreted in a variety of ways in both general education and language learning. In their original formulation of the concept, Glaser and Klaus (1962: 422), in the context of proficiency measurement in military and industrial training, stated that

> *knowledge of an individual's score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do*

Glaser (1963) described criterion-referenced assessment (CRA) thus:

> *The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency. The standard against which a student's performance is compared when measured in this manner is the behaviour which defines each point along the achievement continuum. The term 'criterion', when used this way, does not necessarily refer to final end-of-course behaviour. Criterion levels can be established at any point in instruction as to the adequacy of an individual's performance. The point is that the specific behaviours implied at each level of proficiency can be identified and used to describe the specific tasks a student must be capable of performing before he achieves each of these knowledge levels. It is in this sense that measures of proficiency can be criterion-referenced.*

This early definition of CRA highlights several key elements which are reflected in various kinds of language assessment instruments: first, proficiency (here, interestingly, not distinguished very clearly from achievement) is conceived of as a continuum ranging from no proficiency at all to "perfect" proficiency; second, the *criterion* is defined as an external standard against which learner behaviour is compared; and third, levels of proficiency (or achievement) are linked to specific tasks.

CRITERION-REFERENCING IN LANGUAGE ASSESSMENT

In the context of language learning, CRA has number of different meanings (Skehan 1989: 5-6). In the first instance, it refers in a general sense to tests or assessments which are based on sampling of a behavioural domain and which make explicit the features of this domain. For example, in an oral interview, a testee might be given a score on a rating scale which contains the key aspects of performance (that is, the *criteria*) to be assessed such as fluency, appropriacy, accuracy, pronunciation, grammar etc. These criteria may then be described more fully in a band or level description. As Skehan (1984: 217) notes, such descriptions represent a set of generalised behaviours relating performance to external criteria (referred to by Jones 1985: 82 as the *performance criterion*), rather than a statement that would enable a yes/no decision to be made with respect to a testee's ability on a particular task.

As we have seen, CRA also carries a second meaning of a *standard* (criterion level) or cut-off point which may be defined with reference to some external requirement. In the context of language assessment, this might be exemplified by the "threshold level" set by the Council of Europe as a minimal level of functional language competence. Some writers, in fact, posit the existence of a constant and "natural reference point" for this external standard in the form of the native speaker (see, for example, Cziko 1983: 294).

Skehan (1989) also suggests a third sense in which CRA can be interpreted:

> *This is that the proficiency levels which are the basis for criterion-referencing are linked in some cumulative way to a course of development.*

This raises the issue of whether assessment criteria should take as their reference point what learners do, what linguists and teachers think learners do or what native speakers do. This point will be taken up later.

NORM-REFERENCING VERSUS CRITERION-REFERENCING

CRA is traditionally contrasted with *norm-referenced* methods of assessment which are meant to compare individual's performances relative to each other and to distribute them along the normal curve, not to establish the degree to which students have mastered a particular skill (Hudson and Lynch 1984: 172). Large-scale standardized examinations, in which students are given aggregate scores or grades for purposes of selection, certification or placement are probably the best-known example of norm-referenced assessment. An example of a norm-referenced approach from second language learning would be proficiency test batteries in which results are reported solely in terms of an overall score (a range of such tests is described by Alderson, Krahnke and Stansfield 1987).

According to some authors, however, the differences between norm-referenced assessment and CRA however, are not as great as conventionally imagined. Rowntree (1987: 185-6), for example, notes that criterion levels are frequently established by using population norms:

> *So much assessment that appears to be criterion-referenced is, in a sense, norm-referenced. The difference is that the student's performance is judged and labelled by comparison with the norms established by other students elsewhere rather than those established by his immediate fellow-students.*

There is an element of both norm- and criterion-referencing about the way in which proficiency descriptions are drawn up and interpreted. For example, one method of defining assessment criteria and performance descriptors for writing proficiency is to ask experienced teachers, without the aid of any explicit criteria, to rank learners in order of proficiency by sorting a set of writing scripts into piles representing clearly definable proficiency differences. Following this, the characteristic features of scripts at each level are discussed and these are then used to establish criteria and performance descriptors.

The level descriptions in proficiency scales, as numerous authors have pointed out (eg. Trim 1977, Skehan 1984), often contain norm-referenced terminology despite their claim to be criterion-referenced. Terminology such as "greater flexibility" or "fewer errors" relates the levels to each other instead of to the external standard which is supposed to characterise criterion-referencing. In terms of their actual use, as well, the descriptors may be interpreted in covertly norm-referenced ways. It is not unusual, for example, to hear teachers refer to a "good Level 1", a "slow Level 2" etc.

DEVELOPING CRITERIA AND DESCRIBING PERFORMANCE

Real world and classroom dimensions of CRA

CRA has both a real-world and a classroom dimension. In the development of a proficiency test aimed at assessing real-world language use, defining criteria involves operationalising the construct of proficiency -- in other words, specifying the skills and abilities which constitute the test developer's view of "what it means to know how to use a language" (Spolsky 1986). From the test specifications thus established, items are constructed and/or level/band descriptions written according to which performance will be rated. This is, of necessity, a time-consuming and rigorous process involving empirical studies of performance samples, consultation with expert judges and continuing revision of criteria and descriptors (see, for example, the descriptions by Alderson (1989) and Westaway (1988) of the way in which IELTS bands were derived).

In classroom CRA which is aimed at assessing learner achievement or diagnosing difficulties, the process of defining criteria and descriptors involves specifying the behavioural domain from which objectives are drawn, formulating a set of relevant objectives and establishing a set of standards by which learners' performance is judged. In many ways, this process replicates what is involved in operationalising the construct of proficiency, in that it involves specifying the nature of the domain to be assessed and breaking this down into its component parts. However, classroom CRA is likely to be less formal and may rely on implicit judgements on the teacher's part as to what constitutes the domain of ability which is assessed (Black and Dockrell 1984: 42-43).

It is worth noting at this point that the interpretation of "criterion" is slightly different, according to the purposes for which CRA is being carried out. Where learners' proficiency is being assessed in order to determine their capacity to undertake some real-world activity (eg. to exercise a profession), *criterion-referenced* is often taken to mean that their performance is compared against a "criterion level" of performance or a cut-score. They either reach the criterion or they don't. As Davies (1988: 33) notes, users of tests interpret all test results in a criterion-referenced way. A candidate's actual score is of less importance than the question: has the candidate attained the cut score or not?

In the classroom, however, the emphasis is slightly different. Here, the "criterion" against which learners' performance is assessed relates to a domain specification and a set of learning objectives. Attainment may be assessed in terms of mastery/non-mastery of these objectives (see, for example, Hudson and Lynch 1984; Hudson 1989). However, making a yes/no decision on whether mastery has been attained can be extreme', difficult. In fact, the validity of the concept itself has been questioned (Glass 1978) and there are a multiplicity of competing views on appropriate standard-setting methods in CRA (see Berk 1986 for a comprehensive discussion of the relative merits of various methods). For this reason, classroom CRA is often more concerned with assessing learners' attainment on a scale of ability which represents varying degrees of mastery but is not necessarily linked to a "cut-score" (see Brindley 1989 for examples).

In terms of content, CR proficiency testing tends to focus on assessing tasks which replicate real life or from which inferences can be made to real-life perform. ince. As far as classroom assessment is concerned, however, opinions differ on the question of whether CRA should be exclusively focussed on subsequent extra-classroom tasks or whether *any* valid objective can be assessed (Brown 1981: 7). If the latter view is accepted, then it would be possible to imagine situations in which CRA assessment did not concern itself with elements of learners' communicative performance (eg. if the syllabus were grammatically-based). CRA does not, in other words, necessarily mean communicative assessment. However, in the case of second language learners who have to use the language in society on a daily basis there are clearly arguments for accentuating methods of CRA which allow them to gain feedback on their ability to perform real-life tasks (see Brindley 1989: 91-120 for examples).

DEFINING CRITERIA

A variety of methods have been used by test developers and teachers to define assessment criteria and performance descriptors. These will be described below and some problems associated with each will be discussed.

Use existing criteria

The easiest way to define criteria and descriptors for language assessment is to use those already in existence. There is no shortage of models and examples. For proficiency testing, literally thousands of rating scales, band scales and performance descriptors are used throughout the world. An equivalent number of skills taxonomies, competency checklists, objectives grids etc, are available for classroom use.

Like tests, some proficiency scales seem to have acquired popular validation by virtue of their longevity and extracts from them regularly appear in other scales. The original scale used in conjunction with the Foreign Service Institute Oral Interview (FSI 1968), in particular, seems to have served as a source of inspiration for a wide range of other instruments with a similar purpose but not necessarily with a similar target group. Both the Australian Second Language Proficiency Rating Scale (ASLPR) (Ingram 1984) and the ACTFL Proficiency Guidelines (Hiple 1987) which aim to describe in the first case the proficiency of adult immigrants in Australia and in the second the proficiency of foreign language students and teachers in the USA, draw on the FSI descriptions.

Problems

Although proficiency scales have gained widespread acceptance over a considerable period of time and appear face-valid, it is very difficult to find any explicit information on how the descriptions were actually arrived at. Although some scales are claimed to be data-based (see, for example, Liskin-Gasparro (1984: 37) who states that the ACTFL guidelines were developed empirically), no information is made publicly available as to how the data were collected, analysed and turned into performance descriptors. This is despite the fact that in some cases claims are being made (if only by inference) to the effect that the descriptions constitute universal descriptions of second language development. Byrnes (1987), for example, claims that the ACTFL/ETS scale is built on a "hierarchy of task universals" .

Apart from their lack of empirical underpinning, the validity of rating scale descriptors (in particular the ACTFL/ETS Oral Proficiency Interview) has been contested on a number of other grounds. Some of the principal concerns which have been voiced can be roughly summarised as follows:

- the logic of the way levels are arrived at is essentially circular--"the criteria are the levels and vice versa" (Lantolf and Frawley 1985: 340). They cannot therefore be criterion-referenced in the accepted sense since there is no external standard against which the testee's behaviour may be compared.

- the incremental and lockstep nature of level descriptions fails to take into account the well documented variability and "backsliding" which occur in interlanguage (Pienemann, Johnston and Brindley 1988); nor can differential abilities in different "discourse domains" be accounted for (see Douglas and Selinker 1985, Zuengler 1989). In particular, the assumption that grammatical and phonological accuracy increases in a linear fashion is contradicted by evidence from second language acquisition studies which have shown systematic variability according to the learner's psycho-sociological orientation (Meisel et al. 1981); emotional investment in the topic (Eisenstein and Starbuck 1989); the discourse demands of the task (Brown and Yule 1989); desired degree of social convergence/divergence (Rampton 1987); planning time available (Ellis 1987); and ethnicity and status of interlocutor (Beebe 1983)

- not only are the performance descriptions covertly norm-referenced (see above), but also there is no principled relationship between co-occurring performance features which figure in the one level (Skehan 1984, Brindley 1986).

- it is very difficult to specify relative degrees of mastery of a particular skill with sufficient precision to distinguish clearly between levels. This is illustrated by Alderson's (1989: 11) comment on the development of the IELTS Speaking scales:

  For some criteria, for example pronunciation or grammatical accuracy, the difference in levels came down to a different choice of quantifiers and we were faced with issues like is 'some' more than 'a few' but fewer than 'several' or 'considerable' or 'many'. How many is 'many'?

- the essentially interactive nature of oral communication is inadequately represented due to the restriction of the possible range or roles which can be assumed by the non-native speaker (Lantolf and Frawley 1988; Raffaldini 1988; van Lier 1989).

- the descriptions are highly context dependent and thus do not permit generalisation about underlying ability (Bachman and Savignon 1986; Skehan 1989). Methods such as the oral interview confuse trait and method (Bachman 1988).

- in the absence of concrete upper and lower reference points, criterion-referencing is not possible. Bachman (1989: 17) points out that criterion-referencing requires the definition of the end points of an absolute scale of ability (so-called "zero" and "perfect" proficiency). Yet in practice, no-one has zero proficiency, since some language abilities are universal. Similarly, native speakers vary widely in ability, which makes the "perfect speaker" an equally tenuous concept.

Clearly the validity of the criteria on which proficiency descriptions are built is by no means universally accepted. However, the controversy surrounding the construct validity of proficiency rating scales and performance descriptors is merely a manifestation of the fundamental question that CRA has to face: how to define the domain of ability which is to be assessed, that is, language proficiency? Criterion-referencing depends on a very detailed and exact specification of the behavioural domain. But this amounts to asking the question posed by Spolsky (1986):

What does it mean to know how to use a language?

As far as proficiency testing is concerned, a definitive answer to this question is clearly not presently on the horizon, although detailed and testable models such as that proposed by Bachman (1990) offer some hope of describing more exactly the nature of communicative language ability. Meanwhile, in the context of classroom assessment, the move towards criterion-referencing continues. There is an increasing number of objectives-based assessment and profiling schemes derived from specification of real-life communicative needs which allow cumulative attainment to be monitored and documented in the form of profiles of achievement (see Brindley 1989: 91-111). These present a way of linking classroom assessment closely to real-world outcomes. However, objectives-based domain specifications also require the operationalization of the behaviour which forms the basis of the domain. As such, they are open to question on the same grounds as the proficiency descriptions described above. In addition, some testers would claim that performance testing associated with assessment of course objectives gives no information on underlying ability (Skehan 1989: 7).

The problem of domain specification is clearly far from being resolved. In the meantime, disagreement on the validity of criteria will no doubt continue, since there is as yet no description of language learning and language use on the basis of which universally agreed criteria could be drawn up.

Attacking the domain specification problem

Because of the limitations of context-dependent proficiency descriptions and the difficulties of relating these to an 'absolute' scale of ability, Bachman (1989) argues that the only way to develop adequate CR procedures for assessing communicative language proficiency is to attempt to clearly specify the abilities that make up language proficiency and to define scales or levels of proficiency which are independent of particular contexts, 'in terms of the relative presence or absence of the abilities that constitute the domain' rather than 'in terms of actual individuals or actual performance' (Bachman 1989: 256). An example of such a scale is given below.

	Vocabulary	*Cohesion*
0	*Extremely limited vocabulary*	*No cohesion*
	(A few words and formulaic phrases. Not possible to discuss any topic, due to limited vocabulary).	(Utterances completely disjointed, or discourse too short to judge).
1	*Small vocabulary*	*Very little cohesion*
	(Difficulty in talking with examinee because of vocabulary limitations).	(Relationships between utterances not adequately marked; frequent confusing relationship among ideas)
2	*Vocabulary of moderate size*	*Moderate cohesion*
	(Frequently misses or searches for words).	(Relationships between utterances generally marked; sometimes confusing relationships among ideas).
3	*Large vocabulary*	*Good cohesion*
	(Seldom misses or searches for words).	(Relationship between utterances well-marked).
4	*Extensive vocabulary*	*Excellent cohesion*
	(Rarely, if ever, misses or searches for words. Almost always uses appropriate word)	(Uses a variety of appropriate devices; hardly ever confusing relationships among ideas)

Figure 1 Scales of ability in vocabulary and cohesion (Bachman and Palmer, 1983)

However such scales, too, are clearly fraught with problems as Bachman and Savignon (1986: 388) recognize when they admit the difficulty of 'specifying the degree of control and range in terms that are specific enough to distinguish levels clearly and for raters to interpret consistently'. The sample scales, in fact, manifest many of the same problems which arise in the design of more conventional proficiency rating scales. The terminology used is very imprecise and relativistic ('limited'; 'frequently'; 'confusing' etc) and in the absence of precise examples of learners' language use at each of the levels, problems of rater agreement would inevitably arise. In fact, since the levels do not specify particular contexts, structure, functions and so on, raters would not have any concrete criteria to guide them. The difficulties of reaching agreement between raters would, consequently, be likely to be even more acute.

Consult expert judges

Another commonly used way of producing criteria for proficiency testing is to ask expert judges to identify and sometimes to weight the key features of learner performance which are to be assessed. *Experienced teachers* tend to be the audience most frequently consulted in the development and refining of criteria and performance descriptions (eg. Westaway 1988; Alderson 1989; Griffin 1989). In some cases they may be asked to generate the descriptors themselves by describing key indicators of performance at different levels of proficiency. In others, test developers may solicit comments and suggestions from teachers for modification of existing descriptors on the basis of their knowledge and experience.

In ESP testing, *test users* may also surveyed in order to establish patterns of language usage and difficulty, including the relative importance of language tasks and skills. The survey results then serve as a basis for test specifications. This procedure has been followed in the development of tests of English for academic purposes by, *inter alia*, Powers (1986), Hughes (1988) and Weir (1983, 1988) and by McNamara (1989) in the construction of tests of speaking and writing for overseas-trained health professionals in Australia.

Problems

Who are the experts?

The idea of using "expert judgement" appeals to logic and common sense. However it poses the question of who the experts actually are. Conventionally it is teachers who provide "expert" judgements, although increasingly other non-teacher test users are being involved in test development. There are obvious reasons, of course, for appealing to teacher judgements. They are not difficult to obtain since teachers are on hand, they are familiar with learners' needs and problems, they are able to analyse language and they can usually be assumed to be aware of the purposes and principles of language testing, even though they may not always be sympathetic to it. Although less obviously "expert" in the sense of being further removed from the language learning situation and less familiar with linguistic terminology, test users who interact with the target group (such as staff in tertiary institutions or employers) can similarly be presumed likely to have some idea of the language demands which will be made on the testee and thus to be able to provide usable information for test developers.

But in addition to teachers and test users, it could also be argued that testees/learners themselves are "experts" on matters relating to their own language use and that their perceptions should also be considered in drawing up test criteria and specifications. Self-assessment based on learner-generated criteria is becoming increasingly common practice in classroom-based formative assessment and quite high correlations have been found between self-assessment and other external measures (Oskarsson 1989). However, learner perspectives have only recently begun to figure in proficiency test development (LeBlanc and Painchaud 1985; Bachman and Palmer 1988).

So-called "naive" native speakers constitute another "expert" audience whose perceptions could profitably be drawn on in establishing performance criteria. As Barnwell (1987) forcefully argues:

> *.....the domain of proficiency is outside the classroom not inside. We can (perhaps) leave achievement testing to the teachers and professional testers, but once we aspire to measure proficiency it becomes a question of vox populi, vox dei.*
>
> *Language is central to our humanity, and it is the most democratic and egalitarian attribute we share with our fellow man. Why then should we need 'experts' to tell us how well we speak? Thus it is not just an interesting novelty to contemplate the use of 'native' natives in proficiency testing and rating, it is a logical necessity which arises out of the nature of the thing we are trying to measure.*

Given that it is native speaker judgements of proficiency which may well determine the future of testees, it clearly important to investigate on what basis these judgements are made. As Clark and Lett (1988: 59) point out, comparing native speaker judgements with proficiency descriptors is one way of validating the descriptors in a non-circular way and of establishing the external criteria which have been lacking up to the present.

Data collection is resource-intensive

In order to establish valid performance criteria, an analysis of the testees' future domain of language use is clearly desirable. However, the collection of data for test construction purposes poses a number of logistical difficulties. From a practical point of view, the investigation of communicative needs is extremely resource-intensive, to such an extent that the practical constraints of data-gathering may end up jeopardizing the purpose for which the data are being gathered. (This same point had been made in relation to the rigorous needs assessment procedures which accompanied "target situation analysis" in ESP course development). An example is provided in a study by Stansfield and Powers (1989) aimed at validating the Test of Spoken English as a tool for the selection and certification of non-native health professionals and to establish minimum standards of proficiency. They state:

> *of necessity we asked for relatively global ratings, even for professionals and chose situations that would be representative and typical of those in which each professional might be involved. No attempt was made to specify all the many situations that might be encountered, nor was any effort made to designate highly specific tasks. We might have asked about the degree of speaking proficiency needed in the performance of surgical procedures, for example (in which oral proficiency might be critical) but time limitations precluded such detail. In addition in this study, we decided to consider neither other important dimensions of communicative competence (eg. interpersonal skills and other affective components) nor functions of language (eg. persuading or developing ' ipport with patients) that might be highly desirable in various medical situations.*

In only considering global proficiency, a course of action they were forced to take through lack of necessary resources, the researchers neglected the information which would be considered most essential by some (prospective patients is one group which springs to mind!) for test validity.

Precise information is difficult to elicit

An additional problem in consulting test users or "naive" native speakers in drawing up criteria for assessment is the difficulty of getting them to be sufficiently precise about situations of language use to provide usable information. Powers (1986), reporting on his attempts to elicit information from faculty members on university students' listening patterns, observes that:

> *the notion of analysing listening activities may have been "foreign" to many faculty members who were not involved intensely in language instruction or testing. In particular, such concepts as "discourse cues" and "non-verbal signals" may be somewhat far afielu for faculty in non-language disciplines. Moreover, while the rating of such passive, non-observable skills as listening may be difficult generally, non-language oriented faculty may have even greater difficulty in determining when students encounter specific kinds of problems.*

Native speakers are not language analysts. Nor are most learners. It is hardly surprising, therefore, that the test users' perceptions of language needs tend to be stated in rather vague terms. This is exemplified by an examination by Brindley, Neeson and Woods (1989) of the language-related comments of 63 university supervisors' monitoring reports on the progress of foreign students. They found that the vast majority of the comments were of the general kind ("has problems with writing English"; "English expression not good"), though a few lecturers were able to identify particular goal-related skills ("has difficulty following lecturers-speak very fast").

In a similar vein, Weir (1988: 73), commenting on the development of a test specification framework for the TEEP test of English for academic purposes, notes that

> *There is a need for more precise methods for dealing with task dimensions than the pragmatic ones used in our research. We relied heavily on the judgements of teachers and other experts in the field, as well as on the results of small test administrations, to guide us on the appropriateness of task dimensions in the various constructs. Unless finer instruments are developed than these rather coarse subjective estimates, it is difficult to see how fully parallel versions of the test can ever be developed.*

Expert judgement may be unreliable

If expert opinion is to have any currency as a method of developing criteria, then one would expect that a given group of expert judges would concur, first on the criteria which make up the behavioural domain being assessed and second, on the allocation of particular performance features to particular levels. (Obtaining data in this way would be an integral part of construct validation). One would also expect that the group would be able to agree on the extent to which a test item was testing a particular skill and the level of difficulty represented by the item (agreement would constitute evidence for content validity).

Studies aimed at investigating how expert judgements are made, however, cast some doubt on the ability of expert judges to agree on any of these issues. Alderson (1988), for example, in an examination of item content in EFL reading tests, found that judges were unable to agree not only on what particular items were testing but also on the level of difficulty of items or skills and the assignment of these to a particular level. Devenney (1989) who investigated the evaluative judgements of ESL teachers and students of ESL compositions, found both within-group and between-group differences in the criteria which were used. He comments:

> *Implicit in the notion of interpretive communities are these assumptions: (1) a clear set of shared evaluative criteria exists, and (2) it will be used by members of the interpretive community to respond to text. Yet this did not prove to be the case for either ESL teachers or students*

Different people use different criteria

Non-teacher native speakers, teachers and learners themselves, by virtue of their different backgrounds, experiences and expectations, have different understandings of the nature of language learning and communication. As a result, they tend to use different criteria to judge language ability and thus to pay attention to different features of second language performance. Studies of error gravity, for example, have shown that native speakers tend to be less concerned with grammatical accuracy than teachers (particularly those who are not native speakers of the language taught (Davies 1983)). This highlights the difficulties of constructing assessment criteria and descriptors which can be consistently interpreted by different audiences.

It is interesting, and perhaps significant, to note in the context of this discussion that disciplines outside applied linguistics interpret "communication" or "communicative competence" quite differently and hence employ different criteria for assessment. Communication theorists, for example, accentuate criteria such as *empathy, behavioural flexibility and interaction management* (Wiemann and Backlund 1980) and emphasise the role of non-verbal aspects of communication. In other fields, such as organisational management, communicative ability is seen very much in terms of "getting the job done" and the success of communication is thus judged primarily in relation to how well the outcomes are achieved rather than on specific linguistic features (Brindley 1989: 122-23). McNamara (1987: 32) makes this point in relation to doctor-patient communication, noting that in the medical profession "there is a concern for the communication process in terms of its outcomes". He comments (1987: 47) that "sociolinguistic approaches to 'communicative ability' are indeed narrow, and narrowly concerned with language rather than communicative behaviour as a whole".

Two conclusions can be drawn from these observations. First, as McNamara (op. cit.) points out, we must be conscious of the limitations of the claims which can be made about the capacity of language tests to predict communicative ability (in the broader sense) in real-life settings. Second, if real-life judgements of communicative effectiveness are based on perceptions of people's ability to use language to complete a task satisfactorily, then it is worth trying to build this notion into assessment criteria. In this regard, the use of "task fulfilment" as a criterion in the IELTS writing assessment scales is a promising step in this direction (Westaway 1988).

Teachers will be teachers

Although teachers' judgements are frequently used as a basis for establishing assessment criteria, there is some evidence to suggest that the influence of their background and experience may be sufficiently strong to override the criteria that are given. For example, in a preliminary analysis of 12 videotaped moderation sessions of oral interviews conducted for the purposes of rating speaking ability at class placement in the Australian Adult Migrant Education Program, I have found a consistent tendency for teachers to:

- refer to criteria which are not contained in the performance descriptors at all, such as confidence, motivation, risk-taking capacity and learning potential.

- concentrate heavily on the assessment of some features of performance at the expense of others. In this case, more time was spent discussing the role of the grammatical accuracy than any other single factor, even though the descriptions being used did not provide detailed or specific comments on grammatical features.

- use diagnostically-oriented and judgemental "teacher language" in applying the criteria, such as:

  She seemed to be weak on tenses
  I was a bit concerned about her word order generally

her language was letting her down
She's got weak tense forms, not sure of he: prepositions and quite often leaves off a final-s

Caulley et al (1988), report on a similar phenomenon in the context of the evaluation of common assessment tasks used in the Victorian senior secondary English examination:

> *in their discussions the teachers rarely or even referred to the specified criteria. Their assessments were largely global, the language abstract and rarely substantiated by reference to anything concrete:*

This was exemplified by comments such as

he's got communicative sense
he's more sure of his material
there's a lack of flow
she hasn't crystallised her ideas

They note that

> *teachers are involved with the growth and development of human beings through practice and in the end were shown to be neither willing nor able to divorce the performance of an action from those aspects of it such as intention, effort and risk, which make it one performed by a growing and developing human beings. They thus included in their assessment of students an estimate of the risk involved for the particular student to present as he or she did and something for the effort (or lack of effort) made in the preparation, although neither is mentioned in the guidelines.*

Although such non-linguistic factors do not conventionally figure as criteria in definitions of proficiency, it would appear that they are included by teachers, perhaps because they are perceived as part of their educator's role. Specific assessment criteria may be developed rigorously and clearly spelled out, yet the teachers appear to be operating with their own constructs and applying their own criteria in spite of (or in addition to) those which they are given. This tendency may be quite widespread and seems to be acknowledged by Clark and Grognet

(1985: 103) in the following comment on the external validity of the Basic English Skills Test for non-English-speaking refugees in the USA:

> *On the assumption that the proficiency-rating criterion is probably somewhat unreliable in its own right, as well as based to some extent on factors not directly associated with language proficiency per se (for example, student personality, diligence in completing assignments etc) even higher validity coefficients might be shown using external criteria more directly and accurately reflecting language proficiency*

Further support for the contention that teachers operate with their own criteria is provided by a study carried out by Griffin (1989) who examined the consistency of the rating of IELTS writing scripts over time using a Rasch Rating scale model. An analysis of rater statistics revealed that

> *For assessment 1, most raters appeared to 'fix' the underlying variable. On occasion 2, however, few raters appeared to fix the variable. There appears to have been a change in the criteria or in the nature of the variable being used to assign scripts to levels. The original criteria used in the familiarisation workshop and reinforced in the training workshop do not seem to have been used for assessment 2. Unfortunately it was assumed that the criteria would remain the same and were in fact supplied to the raters.*
>
> (Griffin 1989: 10)

He comments that

> *raters seem to be influenced by their teaching background and the nature of the criteria used can differ from rater to rater. Consensus moderation procedures appear to have controlled this effect to some degree but not completely.*
>
> (Griffin 1989: 13)

CONCLUSION

From this review of CRA, it should be clear, as Skehan (1984: 216) remarks, that "criterion-referencing is an attractive ideal, but extremely difficult to achieve in practice". As we have seen, the criteria which are currently used may not reflect what is known about the nature of language learning and use and they may not be consistently interpreted and applied even by expert judges.

If the ideal of CRA is to be attained, it is necessary to develop criteria and descriptors which not only reflect current theories of language learning and language use but which also attempt to embody multiple perspectives on communicative ability. As far as the first of these requirements is concerned, Bachman and his colleagues have put forward a research agenda to develop operational definitions of constructs in Bachman model of communicative language proficiency and validate these through an extensive program of test development and research (see, for example, Bachman and Clark 1987; Bachman et al 1988; Bachman 1990). One of the main virtues of this model, as Skehan (1990) points out, is that it provides a framework within which language testing research can be organised. It is to be hoped that the model will enable language testers to systematically investigate the components of language ability as manifested in tests and that the results of such research will be used to inform the specifications on which assessment instruments are based.

Second language acquisition (SLA) research can also make a contribution to the development of empirically-derived criteria for language assessment which reflect the inherent variability and intersubjectivity of language use. First, research into *task variability* of the type reported in Tarone (1989), Tarone and Yule (1989) and Gass et al (1989a: 1989b) provides valuable insights into the role that variables such as interlocutor, topic, social status and discourse domain might exercise on proficiency. Investigation of factors affecting *task difficulty* might also provide a more principled basis for assigning tasks to levels, a major problem in CRA. A number of testable hypotheses are outlined by Nunan (1989).

Second, SLA research could also provide much-needed information on the factors which influence native speaker perceptions of non-native speakers' proficiency. There is already a considerable literature on the overall communicative effect of non-native speaker communication (eg Albrechtsen et al 1980; Ludwig 1982; Eisenstein 1983) and error gravity (eg James 1977; Chastain 1980; Davies 1983). However such studies have tended to examine the effects of particular discourse, phonological, syntactic or lexical features on comprehensibility and/or irritation, rather than relating them to perceptions of proficiency. Studies conducted with a specific focus on proficiency would assist in the creation of performance criteria which reflect those used in real life. Information of this kind is of critical importance since in many cases, it is the judgements of native speakers that will determine the future of language learners, not so much those of teachers. At the same time, it is important to try to establish to what extent non-linguistic factors such as personality, social status, ethnicity, gender etc affect judgements of proficiency and the extent to which these factors can be related to linguistic ones (Clark and Lett 1987).

Third, research into the nature of *developmental sequences* in learner language gives an indication of the grammatical elements of language which can realistically be expected for production at different stages and thus provides a basis for establishing assessment criteria which are consistent with the regularities of language development (Pienemann et al 1988). In addition, since the multi-dimensional model of second language acquisition described by Pienemann and Johnston (1987) makes strong predictions concerning the processing demands made by different linguistic elements on learners, it should be possible to incorporate these predictions into concrete hypotheses concerning task difficulty which can be empirically investigated.

Thus far I have sketched out the kinds of research that might contribute to the development of better criteria. As far as the *interpretation* of the criteria is concerned, however, it would be naive to imagine that different judges will not continue to interpret criteria idiosyncratically. As Messick (1989) says:

> *....expert judgement is fallible and may imperfectly apprehend domain structure or inadequately represent test structure or both.*

Agreement between testers can be improved by familiarisation and training sessions in which raters, as Griffin (1989) reports. But there is always the possibility that agreement might conceal fundamental differences. As Barnwell (1985) comments:

> *raters who agree on the level at which a candidate can be placed may offer very different reasons for their decisions*

Given, as we have seen, that different judges may operate with their own personalized constructs irrespective of the criteria they are given, it would be a mistake to assume that high inter-rater reliability constitutes evidence of the construct validity of the scales or performance descriptors that are used. In order to provide such evidence, empirically-based investigation of the behavioural domain itself has to be carried out, as I have indicated above. At the same time, studies requiring teachers, learners and native speakers are to externalize the criteria they (perhaps unconsciously) use to judge language ability would help to throw some light on how judgements are actually made by a variety of different audiences and lead to a better understanding of the constructs that inform the criteria they use. The procedures used in the development of the IELTS band scales as reported by Westaway (1988), Alderson (1989), Griffin (1989) offer the possibility of building up a useful data base in this area.

Finally, in the context of classroom CRA, the time is ripe to explore the feasibility of incorporating communicatively-oriented CRA into the teaching and learning process. In the field of gene^ral education, the results of research into the development of CR instruments for classroom use indicates that the problems of domain specification described in this paper may not be as intractable as they are sometimes portrayed (Black and Dockrell 1984). Numerous CR schemes for formative assessment and profiling are in existence in general education the United Kingdom and Australia (see Brindley 1989 for an overview) and appear to be quite adaptable to second language learning situations. The use of CR methods of assessing achievement based on communicative criteria would not only help to link teaching more closely to assessment, but also would allow for closer involvement of learners in monitoring and assessing their progress.

ACKNOWLEDGEMENT

I would like to thank Charles Alderson for his helpful comments on an earlier version of this paper.

REFERENCES

Alderson, J C. 1989. Bands and scores. Paper presented at IATEFL Language Testing Symposium, Bournemouth, 17-19 November.

Alderson, J C. 1988. Testing reading comprehension skills. Paper presented at the Sixth Colloquium on Research in Reading in a Second Language. TESOL, Chicago, March 1988.

Alderson, J C., K Krahnke and C W Stansfield (Eds.) 1987. Reviews of English Language Proficiency Tests. Washington: TESOL.

Bachman, L F. 1988. Problems in examining the validity of the ACTFL oral proficiency interview. Studies in Second Language Acquisition, 10, 2, pp. 149-164.

Bachman, L F. 1989. The development and use of criterion-referenced tests of language ability in language program evaluation. (In) The Second Language Curriculum, R K Johnson (ed.), Cambridge: Cambridge University Press.

Bachman, L F. 1990. Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Bachman, L F., and S Savignon 1986. The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. Modem Language Joumal, 70, 4, pp. 380-390.

Bachman, L F., and J L D. Clark 1987. The measurement of foreign/second language proficiency. Annals of the American Academy of Political and Social Sciences, 490, pp. 20-33.

Bachman, L F., and A S Palmer 1983. Oral interview test of communicative proficiency in English. MS.

Bachman, L F., and A S Palmer 1988. The construct validation of self-ratings of communicative language ability. Paper presented at Tenth Annual Language Testing Research Colloquium, University of Illinois at Urbana-Champaign, March 5-7.

Beebe, L. 1983. Risk-taking and the language leamer. (In) Classroom-Oriented Research in Second Language Acquisition, H. Seliger and M. Long (Eds.) Rowley, Massachusetts: Newbury House.

Berk, R A. 1986. A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 1, pp. 137-172.

Black, H D and W B Dockrell 1984. Criterion-Referenced Assessment in the Classroom. Edinburgh: Scottish Council for Research in Education.

Brindley, G. 1986. The Assessment of Second Language Proficiency: Issues and Approaches. Adelaide: National Curriculum Resource Centre.

Brindley, G. 1989. Assessing Achievement in the Learner-Centred Curriculum. Sydney: National Centre for English Language Teaching and Research.

Brindley, G S Neeson and S Woods 1989. Evaluation of Indonesia-Australia Language Foundation Assessment Procedures. Unpublished manuscript.

Brown, S. 1981. What do they know? A Review of Criterion-Referenced Assessment. Edinburgh: Her Majesty's Stationery Office.

Byrnes, H. 1987. Proficiency as a framework for research in second language acquisition. Modern Language Journal, 71, 1, pp. 44-49.

Caulley, D., Orton, J., and L Claydon 1988. Evaluation of English oral CAT. Melbourne: Latrobe University.

Chastain, K. 1980. Native speaker reaction to instructor-identified student second language errors. Modern Language Journal, 64, pp. 210-215.

Clark, J L D and J Lett 1987. A research agenda. (In) Second Language Proficiency Assessment: Current Issues, P Lowe and C W Stansfield (Eds.), Englewood Cliffs: Prentice-Hall, pp. 53-82.

Clark, J L D and A Grognet 1985. Development and validation of a performance-based test of ESL 'survival skills". (In) Second Language Performance Testing, P Hauptman, R LeBlanc and M Wesche (Eds.). Ottawa: Ottawa University Press.

Cziko, G. 1983. Psychometric and edumetric approaches to language testing. (In) Issues in Language Testing Research, J W Oller (Ed.), Rowley, Massachusetts: Newbury House, pp. 289-307.

Davies, E. 1983. Error evaluation: the importance of viewpoint. English Language Teaching Journal, 37, 4, 304-311.

Devenney, R. 1989. How ESL teachers and peers evaluate and respond to student writing. RELC Journal, 20, 1, pp. 77-90

Dickinson, L. 1987. Self-Instruction in Language Learning. Cambridge: Cambridge University press.

Douglas, D and L Selinker 1985. Principles for language tests within the 'discourse domains' theory of interlanguage: research, test construction and interpretation. Language Testing, 2, 2, pp. 205-226.

Eisenstein, M. 1983. Native-speaker reactions to non-native speech: a review of empirical research. Studies in Second Language Acquisition, 5, 2, pp. 160-176.

Eisenstein, M and R Starbuck. 1989. The effect of emotional investment on L2 production. (In) Gass et al (Eds.). 1989b.

Ellis, R. 1987. Interlanguage variability in narrative discourse: style-shifting in the use of the past tense. Studies in Second Language Acquisition, 9, 1, pp. 1-20.

Foreign Service Institute 1968. Absolute Language Proficiency Ratings. Washington, D C; Foreign Service Institute.

Gass, S., C Madden, D Preston and L Selinker (Eds) 1989a. Variation in Second Language Acquisition: Discourse and Pragmatics. Clevedon, Avon: Multilingual Matters.

Gass, S., C Madden, D Preston and L Selinker (Eds) 1989b. Variation in Second Language acquisition: Psycholinguistic Issues. Clevedon, Avon: Multilingual Matters.

Glaser, R., and D J Klaus 1962. Assessing human performance. (In) R Gagne (Ed.), Psychological Principles in Systems Development., New York: Holt, Rinehart and Winston.

Glaser, R. 1963. Instructional technology and the measurement of learning outcomes. American Psychologist, 18, pp. 519-521.

Glass, G V. 1978. Standards and criteria. Journal of Educational Measurement, 15, pp. 237-261.

Griffin, P E. 1989. Latent trait estimates of rater relibility in IELTS. Paper presented at Fourteenth Annual Congress of the Applied Linguistics Association of Australia, Melbourne, September 1989.

Hiple, D. 1987. A Progress report on the ACTFL proficiency guidelines 1982-1986. (In) Defining and Developing Proficiency, H Byrnes and M Canale (Eds.), Lincolnwood, Illinois: National Textbook Company.

Hudson, T. 1989. Mastery decisions in program evaluation. (In) The Second Language Curriculum Curriculum, R K Johnson (ed.), Cambridge: Cambridge University Press.

Hudson, T and B Lynch 1984. A criterion-referenced approach to ESL achievement testing. Language Testing, 1, 2, pp. 171-201.

James, C V. 1977. Judgements of error gravity. English Language Teaching Journal, 31, 2, pp. 175-182.

Jones, R L. 1985. Some basic considerations in testing oral proficiency. (In) New Directions in Language Testing, Y P Lee, A C Y Y Fok, R Lord and G Low (eds.), Oxford: Pergamon, pp. 77-84.

Lantolf, J P and W Frawley 1985. Oral proficiency testing: a criticcl analysis. Modem Language Journal, 69, 4.

Lantolf, J P, and W Frawley 1988. Proficiency: understanding the construct. Studies in Second Language Acquisition, 10, 2, pp. 181-195.

LeBlanc, R, and G Painchaud 1985. Self-assessment as a second language placement instrument. TESOL Quarterly, 19, 4, pp. 11-42.

Liskin-Gasparro, J. 1984. The ACTFL guidelines: a historical perspective. (In) Teaching for proficiency: The organizing principle, Lincolnwood, Illinois: National Textbook Company.

McNamara, T F. 1987. Assessing the Language Proficiency of Health Professionals. Recommendations for Reform of the Occupational English Test. Parkville, Victoria, University of Melbourne: Department of Russian and Language Studies.

McNamara, T F. 1989. ESP testing: general and particular. (In) Language, Learning and Community, C N Candlin and T F McNamara (eds.), Sydney: National Centre for English Language Teaching and Research, pp. 125-142.

Meisel, J, H Clahsen and M Pienemann 1981. On determining development stages in second language acquisition. Studies in Second Language Acquisition 3, pp. 109-135.

Messick, S. 1989. Meaning and values in test validation: the science and ethnics of assessment. Educational Researcher, 18, 2, pp. 5-11.

Nunan, D. 1989. Designing Tasks for the Communicative Classroom. Cambridge: Cambridge University Press.

Oskarssen, M. 1984. Self-Assessment of Foreign Language Skills. Strasbourg: Council of Europe.

Oskarsson, M. 1989. Self-assessment of language proficiency: rationale and applications. Language Testing, 6, 1, pp. 247-259

Pienemann, M, and M Johnston 1987. Factors influencing the development of language proficiency. (In) applying Second Language Acquisition Research, D Nunan (ed.), Adelaide: National Curriculum Resource Centre, pp. 45-141.

Pienemann, M., M Johnston and G Brindley 1988. Constructing an acquisition-based assessment procedure. Studies in Second Language Acquisition, 10, 2, pp. 217-243.

Powers, D E. 1986. Academic demands related to listening skills. Language Testing, 3, 1, pp. 1-38.

Powers, D E and C W Stansfield 1989. An approach to the measurement of communicative ability in three health professions. (In) Working with Language, H Coleman (ed.), Berlin: Mouton de Gruyter, pp 341-366.

Raffaldini, T. 1988. The use of situation tests as measures of communicative ability. Studies in Second Language Acquisition, 10, 2, pp. 197-216.

Rampton, B. 1987. Stylistic variability and not speaking 'normal' English: some post-Labovian approaches and their implications for the study of interlanguage. (In) Second Language Acquisition in Context. R Ellis (Ed.), Englewood Cliffs: Prentice Hall, pp. 47-58.

Rowntree, D. 1977. Assessing Students: How Shall We Know Them? London: Harper and Row.

Skehan, P. 1984. Issues in the testing of English for specific purposes. Language Testing, 1, 2, pp. 202-220.

Skehan, P. 1988. State of the art: language testing. Part 1. Language Teaching, 21, 2, pp. 211-221.

Skehan, P. 1989. State of the art: language testing. Part 2. Language Teaching, 22, 1, pp. 1-13.

Skehan, P. 1990. Progress in language testing: the 1990s. Revised version of plenary address to IATEFL Language Testing Symposium, Bournemouth, 17-19 November 1989.

Spolsky, B. 1985. What does it mean to know how to use a language? An essay on the theoretical basis of language testing. Language Testing, 2, 2, pp. 180-191.

Tarone, E and G Yule 1989. *Focus on the Language Learner.* Oxford: Oxford University Press.

Tarone, E. 1988. *Variation in Interlanguage.* London: Edward Arnold.

Trim, J L M. 1977. *Some Possible Lines of Development for an Overall Structure for a European Unit/Credit Scheme for Foreign Language Learning by Adults.* Strasbourg: Council of Europe.

van Lier, L. 1988. Reeling, writhing, fainting and stretching in coils: oral proficiency interviews as conversation. *TESOL Quarterly*, pp. 489-508.

Westaway, G. 1988. Developments in the English Language Testing Service (ELTS) M2 writing test. *Australian Review of Applied Linguistics*, 11, 2, pp. 13-29.

Wiemann, J M., and P Backlund 1980. Current theory and research in communicative competence. *Review of Educational Research*, 50, 1, pp. 185-199.

Wier, C J. 1988. The specification, realization and validation of an English language proficiency test. (In) *Testing English for University Study*, A Hughes (Ed.), London: Modern English Publications and the British Council.

Zuengler, J. 1989. Performance variation in NS-NNS interactions: ethnolinguistics difference or discourse domain? (In) Gass et al (Eds) 1989a pp. 228-243.

THE ROLE OF ITEM RESPONSE THEORY IN LANGUAGE TEST VALIDATION

T F McNamara

INTRODUCTION

The last decade has seen increasing use of Item Response theory in the examination of the qualities of language tests. Although it has sometimes been seen exclusively as a tool for improved investigation of the *reliability* of tests (Skehan, 1989), its potential for investigation of aspects of the *validity* of language tests has also been demonstrated (McNamara, 1990). However, the application of IRT in this latter role has in some cases met with objections based on what are claimed to be the unsatisfactory theoretical assumptions of IRT, in particular the so-called 'unidimensionality' assumption (Hamp-Lyons, 1989). In this paper, these issues will be discussed in the context of the analysis of data from an ESP Listening test for health professionals, part of a larger test, the Occupational English Test (OET), recently developed on behalf of the Australian Government (McNamara, 1989b).

The paper is in three sections. First, there is a brief description of the Listening sub-test of the OET. Second, the appropriateness of the use of IRT in language testing research is discussed. Third, the use of IRT in the validation of the Listening sub-test of the OET is reported. In this part of the paper, the issue of unidimensionality is considered in the context of analysis of data from the two parts of this test.

THE LISTENING SUB-TEST OF THE OCCUPATIONAL ENGLISH TEST

The Occupational English Test (McNamara, 1989b) is administered to several hundred immigrant and refugee health professionals wishing to take up practice in Australia each year. The majority of these are medical practitioners, but the following professional groups are also represented: nurses, physiotherapists, occupational therapists, dentists, speech pathologists and veterinary surgeons, among others. Responsibility for administering the test lies with the National Office for Overseas Skills Recognition (NOOSR), part of the Commonwealth

Government's Department of Employment, Education and Training. NOOSR was established in 1989 as an expanded version of what had been until then the Council for Overseas Professional Qualifications (COPQ).

The OET is taken as one of three stages of the process of registration for practice in Australia (the other stages involve pencil-and-paper and practical assessments of relevant clinical knowledge and skills). Prior to 1987, the OET was a test of general English proficiency and was attracting increasing criticism from test takers and test users in terms of its validity and reliability. In response to this, COPQ initiated a series of consultancies on reform of the test. The report on the first of these, which was carried out by a team at Lancaster University, recommended the creation of a test which would (Alderson et al., 1986: 3)

assess the ability of candidates to communicate effectively in the workplace.

A series of further consultancies (McNamara, 1987 ; McNamara, 1988a; McNamara, 1989a) established the form of the new test and developed and trialled materials for it. There are four sub-test, one each for Speaking, Listening, Reading and Writing. The format of the new test is described in McNamara (1989b). The validation of the Speaking and Writing sub-tests is discussed in McNamara (1990).

The Listening sub-test is a 50-minute test in two parts. Part A involves listening to a talk on a professionally relevant subject. There are approximately twelve short answer questions, some with several parts; the maximum score on this part of the test is usually about twenty-five. Part B involves listening to a consultation between a general practitioner and a patient. There are approximately twenty short answer questions (again, some have several parts); the maximum score here is usually twenty-five, giving a total maximum score of approximately fifty on thirty-two items. Because of test security considerations, new materials are developed for each session of the test, which is held twice a year.

Before going on to report on the use of IRT in the validation of the Listening sub-test of the OET, the debate about the appropriateness of the use of IRT in language testing research will be reviewed.

Applications of IRT in language testing

The application of IRT to the area of language testing is relatively recent. Oller (1983) contains no reference to IRT in a wide-ranging collection. By contrast, IRT has featured in a number of studies since the early 1980s. Much of this work has focused on the advantages of IRT over classical theory in investigating the reliability of tests (eg Henning, 1984). More significant is the use of IRT to examine aspects of the validity, in particular the construct validity, of tests.

de Jong and Glas (1987) examined the construct validity of tests of foreign language listening comprehension by comparing the performance of native and non-native speakers on the tests. It was hypothesized in this work that native speakers would have a greater chance of scoring right answers on items: this was largely borne out by the data. Moreover, items identified in the analysis as showing 'misfit' should not show these same properties in relation to native speaker performance as items not showing misfit (that is, on 'misfitting' items native speaker performance will show greater overlap with the performance of non-native speakers); this was also confirmed. The researchers conclude (de Jong and Glas, 1987: 191):

> *The ability to evaluate a given fragment of discourse in order to understand what someone is meaning to say cannot be measured along the same dimension as the ability to understand aurally perceived text at the literal level. Items requiring literal understanding discriminate better between native speakers and non-native learners of a language and are therefore better measures of foreign language listening comprehension.*

This finding is provocative, as it seems to go against current views on the role of inferencing processes and reader/listener schemata is comprehension (cf. Carrell, Devine and Eskey, 1988; Widdowson, 1983; Nunan 1987a). One might argue that the IRT analysis has simply confirmed the erroneous assumption that the essential construct requiring measurement is whatever distinguishes the listening abilities of native- and non-native speakers. An alternative viewpoint is that there will in fact be considerable overlap between the abilities of native- and non-native speakers in higher-level cognitive tasks involved in discourse comprehension. If the analysis of listening test data reveals that all test items fail to lie on a single dimension of listening ability, then this is in itself a valid finding about the multi-dimensional nature of listening comprehension in a foreign language and should not be discounted. The point is that interpretation of the results of IRT analysis must be informed by an *in principle* understanding of the relevant constructs.

In the area of speaking, the use of IRT analysis in the development of the Interview Test of English as a Second Language (ITESL) is reported in Adams, Griffin and Martin, 1987; Griffin, Adams, Martin and Tomlinson, 1988. These authors argue that their research confirms the existence of a hypothesized 'developmental dimension of grammatical competence... in English S[econd] L[anguage] A[cquisition]' (1988: 12). This finding has provoked considerable controversy. Spolsky (1988: 123), in a generally highly favourable review of the test, urges some caution in relation to the claims for its construct validity:

> *The authors use their results to argue for the existence of a grammatical proficiency dimension, but some of the items are somewhat more general. The nouns, verbs and adjectives items for instance are more usually classified as vocabulary. One would have liked to see different kinds of items added until the procedure showed that the limit of the unidimensionality criterion had now been reached.*

Nunan (1988: 56) is quite critical of the test's construct validity, particularly in the light of current research in second language acquisition:

> *The major problem that I have with the test...[is] that it fails adequately to reflect the realities and complexities of language development.*

Elsewhere, Nunan (1987b: 156) is more trenchant:

> *[The test] illustrates quite nicely the dangers of attempting to generate models of second language acquisition by running theoretically unmotivated data from poorly conceptualized tests through a powerful statistical programme.*

Griffin has responded to these criticisms (cf Griffin, 1988 and the discussion in Nunan, 1988). However, more recently, Hamp-Lyons (1989) has added her voice to the criticism of the ITESL. She summarizes her response to the study by Adams, Griffin and Martin (1987) as follows (1989: 117):

> *..This study... is a backward step for both language testing and language teaching.*

She takes the writers to task for failing to characterize properly the dimension of 'grammatical competence' which the study claims to have validated; like Spolsky and Nunan, she finds the inclusion of some content areas puzzling in such a test. She argues against the logic of the design of the research project (1989: 115):

> *Their assumption that if the data fit the psychometric model they de facto validate the model of separable grammatical competence is questionable. If you construct a test to test a single dimension and then find that it does indeed test a single dimension, how can you conclude that this dimension exists independently of other language variables? The unidimensionality, if that is really what it is, is an artifact of the test development.*

On the question of the unidimensionality assumption, Hamp-Lyons (1989: 114) warns the developers of the ITESL test that they have a responsibility to acknowledge

> *...the limitations of the partial credit model, especially the question of the unidimensionality assumption of the partial credit model, the conditions under which that assumption can be said to be violated, and the significance of this for the psycholinguistic questions they are investigating... They need to note that the model is very robust to violations of unidimensionality.*

She further (1989: 116) criticizes the developers of the ITESL for their failure to consider the implications of the results of their test development project for the classroom and the curriculum from which it grew.

Hamp-Lyons's anxieties about the homogeneity of items included in the test, echoed by Nunan and Spolsky, seem well-founded. But this is perhaps simply a question of revision of the test content. More substantially, her point about the responsibilities of test developers to consider the backwash effects of their test instruments is well taken, although some practical uses of the test seem unexceptionable (for example, as part of a placement procedure; cf the discussion reported in McNamara, 1988b: 57-61). Its diagnostic function is perhaps more limited, though again this could probably be improved by revision of the test content (although for a counter view on the feasibility of diagnostic tests of grammar, see Hughes, 1989: 13-14).

However, when Adams, Griffin and Martin (1987: 25) refer to using information derived from the test

> *in monitoring and developing profiles,*

they may be claiming a greater role for the test in the curriculum. If so, this requires justification on a quite different basis, as Hamp-Lyons is right to point out. Again, a *priori* arguments about the proper relationship between testing and teaching must accompany discussion of research findings based on IRT analysis.

A more important issue for this paper is Hamp-Lyons's argument about the unidimensionality assumption. Here it seems that she may have misinterpreted the claims of the model, which *hypothesizes* (but does not assume in the sense of 'take for granted' or 'require') a single dimension of ability and difficulty. Its analysis of test data represents a test of this hypothesis in relation to the data. The function of the fit t-statistics, a feature of IRT analysis, is to indicate the probability of a particular pattern of responses (to an item or on the part of an individual) in the case that this hypothesis is true. Extreme values of t, particularly extreme positive values of t, are an indication that the hypothesis is unlikely to be true for the term or the individual concerned. If items or individuals are found in this way to be disconfirming the hypothesis, this may be interpreted in a number of ways. In relation to items, it may indicate (1) that the item is poorly constructed; (2) that if the item is well-constructed, it does not form part of the same dimension as defined by other items in the test, and is therefore measuring a different construct or trait. In relation to persons, it may indicate (1) that the performance on a particular item was not indicative of the candidate's ability in general, and may have been the result of irrelevant factors such as fatigue, inattention, failure to take the test item seriously, factors which Henning (1987: 96) groups under the heading of *response validity*; (2) that the ability of the candidates involved cannot be measured appropriately by the test instrument, that the pattern of responses cannot be explained in the same terms as applied to other candidates, that is, there is a heterogeneous test population in terms of the hypothesis under consideration; (3) that there may be surprising gaps in the candidate's knowledge of the areas covered by the test; this information can then be used for diagnostic and remedial purposes.

A further point to note is that the dimension so defined is a *measurement* dimension which is constructed by the analysis, which must be distinguished from the dimensions of underlying knowledge or ability which may be hypothesized on other, theoretical grounds. IRT analyses do not 'discover' or 'reveal' existing underlying dimensions, but rather construct dimensions for the purposes of measurement on the basis of test performance. The relationship between these two conceptions of dimensionality will be discussed further below.

Hamp-Lyons is in effect arguing, then, that IRT analysis is insufficiently sensitive in its ability to detect in the data departures from its hypothesis about an underlying ability-difficulty continuum. The evidence for this claim, she argues, is in a paper by Henning, Hudson and Turner (1985), in which the appropriateness of Rasch analysis with its attempt to construct a single dimension is questioned in the light of the fact that in language test data (Henning, Hudson and Turner, 1985: 142)

> ...examinee performance is confounded with many cognitive and affective test factors such as test wiseness, cognitive style, test-taking strategy, fatigue, motivation and anxiety. Thus, no test can strictly be said to measures one and only one trait.

(In passing, it should be noted that these are not the usual grounds for objection to the supposedly unidimensional nature of performance on language tests, as these factors have been usefully grouped together elsewhere by Henning under the heading of *response validity* (cf above). The more usual argument is that the *linguistic* and *cognitive skills* underlying performance on language tests cannot be conceptualized as being of one type.) Henning et al. examined performance of some three hundred candidates on the UCLA English as a Second Language

Placement Examination. There were 150 multiple choice items, thirty in each of five sub-tests: Listening Comprehension, Reading Comprehension, Grammar Accuracy, Vocabulary Recognition and Writing Error Detection. Relatively few details of each sub-test are provided, although we might conclude that the first two sub-tests focus on language use and the other three on language usage. This assumes that inferencing is required to answer questions in the first two sub-tests; it is of course quite possible that the questions mostly involve processing of literal meaning only, and in that sense to be rather more like the other sub-tests (cf the discussion of this point in relation to de Jong and Glas (1987) above). The data were analysed using the Rasch one-parameter model, and although this is not reported in detail, it is clear from Table two on p. 153 that eleven misfitting items were found, with the distribution over the sub-tests as follows: Listening, 4; Reading, 4; Grammar, 1; Vocabulary, 3; Writing error detection, 3. (Interestingly, the highest numbers of misfitting items were in the Listening and Reading sub-test). One might reasonably conclude that the majority of test items may be used to construct a single continuum of ability and difficulty. We must say 'the majority' because in fact the Rasch analysis does identify a number of items as not contributing to the definition of a single underlying continuum; unfortunately, no analysis is offered of these items, so we are unable to conclude whether they fall into the category of poorly written items or into the category of sound items which define some different kind of ability. It is not clear what this continuum should be called; as stated above, investigation of what is required to answer the items, particularly in the Reading and Listening comprehension sub-test, is needed. In order to gain independent evidence for the Rasch finding of the existence of a single dimension underlying performance on the majority of items in the test, Henning et al. report two other findings. First, factor analytic studies on previous versions of the test showed that the test as a whole demonstrated a single factor solution. Secondly, the application of a technique known as the Bejar technique for exploring the dimensionality of the test battery appeared to confirm the Rasch analysis findings. Subsequently, Henning et al.'s use of the Bejar technique has convincingly been shown to have been unrevealing (Spurling, 1987a; Spurling, 1987b). Henning et al. nevertheless conclude that the fact that a single dimension of ability and difficulty was defined by the Rasch analysis of their data despite the apparent diversity of the language subskills included in the tests shows that Rasch analysis is (Henning, Hudson and Turner, 1985: 152)

> sufficiently robust with regard to the assumption of unidimensionality to permit applications to the development and analysis of language tests.

(Note again in passing that the analysis by this point in the study is examining a rather different aspect of the possible inappropriateness or otherwise of IRT in relation to language test data than that proposed earlier in the study, although now closer to the usual grounds for dispute). The problem here, as Hamp-Lyons is right to point out, is that what Henning et al. call 'robustness' and take to be virtue leads to conclusions which, looked at from another point of view, seem worrying. That is, the unidimensional construct defined by the test analysis seems in some sense to be at odds with the *a priori* construct *validity*, or at least the face validity, of the test being analysed, and at the very least needs further discussion. However, as has been shown above, the results of the IRT analysis in the Henning study are ambiguous, the nature of the tests being analysed is not clear, and the definition of a single construct is plausible on one reading of the sub-tests' content. Clearly, as the results of the de Jong and Glass study show (and whether or not we agree with their interpretation of those results), IRT analysis is capable of defining different dimensions of ability within a test of a single language sub-skill, and is not necessarily 'robust' in that sense at all, that is, the sense that troubles Hamp-Lyons.

In a follow-up study, Henning (1988: 95) found that fit statistics for both items and persons were sensitive to whether they were calculated in unidimensional or multidimensional contexts, that is, they were sensitive to 'violations of unidimensionality'. (In this study, multidimensionality in the data was confirmed by factor analysis.) However, it is not clear why fit statistics should have been used in this study; the measurement model's primary claims are about the estimates of person ability and item difficulty, and it is these estimates which should form the basis of argumentation (cf the advice on this point in relation to item estimates in Wright and Masters, 1982: 114-117).

In fact, the discussions of Hamp-Lyons and Henning are each marked by a failure to distinguish two types of model: a measurement model and a model of the various skills and abilities potentially underlying test performance. These are not at all the same thing. The measurement model posited and tested by IRT analysis deals with the question, 'Does it mal.: sense in measurement terms to sum scores on different parts of the test? Can all items be summed meaningfully? Are all candidates being measured in the same terms?' This is the 'unidimensionality' assumption; the alternative position requires us to say that separate, qualitative statements about performance on each test item, and of each candidate, are the only valid basis for reporting test performance. All tests which involve the summing of scores across different items or different test parts make the same assumption. It should be pointed out, for example, that classical item analysis makes the same 'assumption' of unidimensionality, but lacks tests of this 'assu:nption' to signal violations of it. As for the interpretation of test scores, this must be done in the light of the our best understanding of the nature of language abilities, that is, in the light of current models of the constructs models such as IRT, and both kinds of analysis have the potential to illuminate the nature of what is being measured in a particular language test.

It seems, then, that Hamp-Lyons's criticisms of IRT on the score of unidimensionality are unwarranted, although, as stated above, results always need to be interpreted in the light of independent theoretical perspective. In fact, independent evidence (of example via factor analysis) may be sought for the conclusions of an IRT analysis when there are grounds for doubting them, for example when they appear to overturn long- or dearly-held beliefs about the nature of aspects of language proficiency. Also, without wishing to enter into Hamp-Lyons (1989: 11[illegible]) calls

> *the hoary issue of whether language competence is unitary or divisible,*

it is clear that there is likely to be a degree of commonality or shared variance on tests of language proficiency of various types, particularly at advanced levels (cf the discussions in Henning (1989: 98) and de Jong and Henning (1990) of recent evidence in relation to this point).

Hamp-Lyons (1989) contrasts Griffin et al.'s work on the ITESL with a study on writing development by Pollitt and Hutchinson (1987), whose approach she views in a wholly positive light. Analysis of data from performance by children in the middle years of secondary school on a series of writing tasks in English, their mother tongue in most cases, led to the following finding (Pollitt and Hutchinson, 1987: 88):

> *Different writing tasks make different demands, calling on different language functions and setting criteria for competence that are more or less easy to meet.*

Pollitt (in press, quoted in Skehan, 1989: 4)

> *discusses how the scale of difficulty identified by IRT can be related to underlying cognitive stages in the development of a skill.*

For Hamp-Lyons (1989: 113), Pollitt and Hutchinson's work is also significant as an example of a valuable fusion of practical test development and theory building.

Several other studies exist which use the IRT Rating Scale model (Andrich, 1978a; Andrich, 1978b; cf Wright and Masters, 1982) to investigate assessments of writing (Henning and Davidson, 1987; McNamara, 1990), speaking (McNamara, 1990) and student self assessment of a range of language skills (Davidson and Henning, 1985). These will not be considered in detail here, but demonstrate further the potential of IRT to investigate the validity of language assessments.

THE OET LISTENING SUB-TEST: DATA

Data from 196 candidates who took the Listening sub-test in August, 1987 were available for analysis using the Partial Credit Model (Wright and Masters, 1982) with the help of facilities provided by the Australian Council for Education Research. The material used in the test had been trialled and subsequently revised prior to its use in the full session of the OET. Part A of the test consisted of short answer questions on a talk about communication between different groups of health professionals in hospital settings. Part B of the test involved a guided history taking in note form based on a recording of a consultation between a doctor and a patient suffering headaches subsequent to a serious car accident two years previously. Full details of the materials and the trialling of the test can be found in McNamara (in preparation).

The analysis was used to answer the following question:

1. Is it possible to construct a single measurement dimension of 'listening ability' from the data from the test as a whole? Does it make sense to add the scores from the two parts of the Listening sub-test? That is, is the Listening test 'unidimensional'?

2. If the answer to the first question is in the affirmative, can we distinguish the skills involved in the two Parts of the sub-test, or are essentially the same skills involved in both? That is, what does the test tell us about the nature of the listening skills being tapped in the two parts of the sub-test? And from a practical point of view, if both sub-tests measure the same skills, could one part of the sub-test be eliminated in the interests of efficiency?

Two sorts of evidence were available in relation to the first question. Candidates' responses were analysed twice. In the first analysis, data from Parts A and B were combined, and estimates of item difficulty and person ability were calculated. Information about departures from unidimensionality were available in the usual form of information about 'misfitting' items and persons. In the second analysis, Part A and Part B were each treated as separate tests, and estimates of item difficulty and person ability were made on the basis of each test separately. It follows that if the Listening sub-test as a whole is unidimensional, then the estimates of person ability from the two separate Parts should be identical; that is, estimates of person ability should be independent of the part of the test on which that estimate is based. The analysis was carried out using the programme MSTEPS (Wright, Congdon and Rossner, 1987).

Using the data from both parts as a single data set, two candidates who got perfect scores were excluded from the analysis, leaving data from 194 candidates. There were a maximum of forty-nine score points from the thirty-two items. Using data from Part A only, scores from five candidates who got perfect scores or scores of zero were excluded, leaving data from 191 candidates. There were a maximum of twenty-four score points from twelve items. Using data from Part B only, scores of nineteen candidates with perfect scores were excluded, leaving data from 177 candidates. There were a maximum of twenty-five score points from twenty items. Table 1 gives summary statistics from each analysis. The *Test reliability of person separation* (the proportion of the observed variance in logit measurements of ability which is not due to measurement error; Wright and Masters, 1982: 105-106), termed the 'Rasch analogue of the familiar KR20 index' by Pollitt and Hutchinson (1987: 82), is higher for the test as a whole than for either of the two parts treated independently. The figure for the test as a whole is satisfactory (.85).

Table 1 Summary statistics, Listening sub-test

	Parts A and B	Part A	Part B
N	194	191	177
Number of items	32	12	20
Maximum raw score	49	24	25
Mean raw score	34.2	14.4	19.4
S D (raw scores)	9.5	5.3	4.5
Mean logit score	1.46	0.86	1.67
S D (logits)	1.33	1.44	1.25
Mean error (logits)	.48	.71	.75
Person separation reliability (like KR-20)	.85	.74	.60

Table 2 gives information on misfitting persons and items in each analysis.

Table 2 Numbers of misfitting items and persons, Listening sub-test

	Parts A and B	Part A	Part B
Items	2 (#7, #12)	2 (#7, #12)	1 (#25)
Persons	2	1	5

The analysis reveals that number of misfitting items is low. The same is true for misfitting persons, particularly for the test as a whole and Part A considered independently. Pollitt and Hutchinson (1987: 82) point out that we would normally expect around 2% of candidates to generate fit values above +2.

On this analysis, then, it seems that when the test data are treated as single test, the item and person fit statistics indicate that all the items except two combine to define a single measurement dimension; and the overwhelming majority of candidates can be measured meaningfully in terms of the dimension of ability so constructed. Our first question has been answered in the affirmative.

It follows that if the Listening sub-test as a whole satisfies the unidimensionality assumption, then person ability estimates derived from each of the two parts of the sub-test treated separately should be independent of the Part of the test on which they are made. Two statistical tests were used for this purpose.

The first test was used to investigate the research hypothesis of a perfect correlation between the ability estimates arrived at separately by treating the data from Part A of the test independently of the data from Part B of the test. The correlation between the two sets of ability estimates was calculated, corrected for attenuation by taking into account the observed reliability of the two parts of the test (Part A: .74, Part B: .60 - cf Table 1 above). (The procedure used and its justification are explained in Henning, 1987: 85-86.) Let the ability estimate of Person n on Part A of the test be denoted by bnA and the ability estimate of Person n on Part B of the test be denoted by bnB. The correlation between these two ability estimates, uncorrected for attenuation, was found to be .74. In order to correct for attenuation, we use the formula

$$rxy = \frac{Rxy}{\sqrt{rxx\ ryy}}$$

where rxy = the correlation corrected for attenuation
Rxy = the observed correlation, uncorrected
rxx = the reliability coefficient for the measure of the variable x
ryy = the reliability coefficient for the measure of the variable y

and where if $rxy > 1$, report $rxy = 1$.

The correlation thus corrected for attenuation was found to be > 1, and hence may be reported as 1. This test, then, enables us to reject the hypothesis that there is not a perfect linear relationship between the ability estimates from

each part of the test, and thus offers support for the research hypothesis that the true correlation is 1.

The correlation test is only a test of the *linearity* of the relationship between the estimates. As a more rigorous test of the *equality* of the ability estimates, a X^2 test was done. Let the 'true' ability of person n be denoted by ßn. Then b*nA* and b*nB* are estimates of ßn. It follows from maximum likelihood estimation theory (Cramer, 1946) that, because bnA and bnB are maximum likelihood estimators of ßn (in the case when both sets of estimates are centred about a mean of zero),

$$bnA \sim N\ (ßn, e^2nA)$$

where *enA* is the error of the estimate of the estimate of the ability of Person n on Part A of the test and

$$bnB \sim N\ (ßn1, e^2nB)$$

where *enB* is the error of the estimate of the ability of Person n on Part B of the test.

From Table 1, the mean logit score on Part B of the test is 1.67, while the mean logit score on Part A of the test is .86. As the mean ability estimates for the scores on each part of the test have thus not been set at zero (due to the fact that items, not people, have been centred), allowance must be made for the relative difficulty of each part of the test (Part B was considerably less difficult than Part A). On average, then, bnB - bnA = .81. It follows that if the hypothesis that the estimates of ability from the two parts of the test are identical is true, then bnB - bnA - .81 = 0. It also follows from above that

$$bnB - bnA - .81 \sim N(0, \overset{2}{e}nB + \overset{2}{e}nA)$$

and thus that

$$\frac{bnB - bnA - .81}{\sqrt{\overset{2}{e}nB + \overset{2}{a}nA}} \sim N\ (0,1)$$

if the differences between the ability estimates (corrected for the relative difficulty of the two parts of the test) are converted to z-scores, as in the above formula. If the hypothesis under consideration is true, then the resulting set of z-scores will have a unit normal distribution; a normal probability plot of these z-scores can be done to confirm the assumption of normality. These z-scores for each candidate are then squared to get a value of X^2 for each candidate. In order to evaluate the hypothesis under consideration for the entire set of scores, then the test statistic is

$$X^2_{N-1} = \sum_{i=1}^{N} z^2$$

where N = 174

The resulting value of X^2 is 155.48,*df* = 173, *p* = .84. (The normal probability plot confirmed that the z-scores were distributed normally). The second statistical test thus enables us to reject the hypothesis that the ability estimates on the two parts of the test are not identical, and thus offers support for the research hypothesis of equality.

The two statistical tests thus provide strong evidence for the assumption of unidimensionality in relation to the test as a whole, and confirm the findings of the analysis of the data from the whole test taken as a single data set. In contrast to the previously mentioned study of Henning (1988), which relied on an analysis of fit statistics, the tests chosen are appropriate, as they depend on ability estimates directly.

Now that the unidimensionality of the test has been confirmed, performance on items on each part of the test may be considered. Figure 1 is a map of the difficulty of items using the data from performance on the test as a whole (N = 194).

Figure 1 Item difficulty map

Difficulty	Item
5.0	
4.0	8
3.0	
2.0	2 3 5 29
	1 12
	11
1.0	25
	15
	7
	16 24
0.0	26
	10 22 23
	6 14 17
	9 18 32
	20 21
-1.0	27 30
	13 28
	4 19
-2.0	
	31
-3.0	

Figure 1 reveals that the two Parts of the test occupy different areas of the map, with some overlap. For example, of the eight most difficult items, seven are from Part A of the test (Part A contains twelve items); conversely, of the eight easiest items, seven are from Part B of the test (Part B has twenty items). It is clear then that differing areas of ability are tapped by the two parts of the test. This is most probably a question of the content of each part; Part A involves following an abstract discourse, whereas Part B involves understanding details of concrete events and personal circumstances in the case history. The two types of listening task can be viewed perhaps in terms of the continua *more or less cognitively demanding and more or less context embedded* proposed by Cummins (1984). The data from the test may be seen as offering support for a similar distinction in the context of listening tasks facing health professionals working through the medium of a second language. The data also offer evidence in support of the content validity of the test, and suggest that the two parts are sufficiently distinct to warrant keeping both. Certainly, in terms of backwash effect, one would not want to remove the part of the test which focuses on the consultation, as face-to face communication with patients is perceived by former test candidates as the most frequent and the most complex of the communication tasks facing them in clinical settings (McNamara, 1989b).

The interpretation offered above is similar in kind to that offered by Pollitt and Hutchinson (1987) of task separation in a test of writing, and further illustrates the potential of IRT for the investigation of issues of validity as well as reliability in language tests (McNamara, 1990).

CONCLUSION

An IRT Partial Credit analysis of a two-part ESP listening test for health professionals has been used in this study to investigate the controversial issue of test unidimensionality, as well as the nature of listening tasks in the test. The analysis involves the use of two independent tests of unidimensionality, and both confirm the finding of the usual analysis of the test data in this case, that is, that it is possible to construct a single dimension using the items on the test for the measurement of listening ability in health professional contexts. This independent confirmation, together with the discussion of the real nature of the issues involved, suggest that the misgivings sometimes voiced about the limitations or indeed the inappropriateness of IRT for the analysis of language test data may not be justified. This is not to suggest, of course, that we should be uncritical of applications of the techniques of IRT analysis.

Moreover, the analysis has shown that the kinds of listening tasks presented to candidates in the two parts of the test represent significantly different tasks in terms of the level of ability required to deal successfully with them. This further confirms the useful role of IRT in the investigation of the content and construct validity of language tests.

REFERENCES

Adams, R J, P E Griffin and L Martin (1987). A latent trait method for measuring a dimension in second language proficiency. Language Testing 4,1: 9-27

Alderson, J C, C N Candlin, C M Clapham, D J Martin and C J Weir (1986). Language proficiency testing for migrant professionals: new directions for the Occupational English Test University of Lancaster.

Andrich, D (1978a). A rating formulation for ordered response categories. Psychometrika 43: 561-573.

Andrich, D (1978b). Scaling attitude items constructed and scored in the Likert tradition. Educational and Psychological Measurement 38: 665-680.

Carrell, P L, J Devine and D E Eskey (eds) (1988). Interactive approaches to second language reading. Cambridge: Cambridge University Press.

Cramer, H (1946). Mathematical methods of statistics. Princeton: Princeton University Press.

Cummins, J (1984). Wanted: a theoretical framework for relating language proficiency to academic achievement among bilingual studies. In C Rivera (ed.) Language proficiency and academic achievement. Clevedon, Avon: Multilingual Matters, 2-19.

Davidson, F and G Henning (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. Language Testing 2,2: 164-179.

De Jong, J.H.A.L and C.A.W Glas (1987). Validation of listening comprehension tests using item response theory. Language Testing 4,2: 170-194.

De Jong, J.H.A.L and G Henning (1990). Testing dimensionality in relation to student proficiency. Paper presented at the Language Testing Research Colloquium, San Francisco, March 2-5.

Griffin P (1988). Tests must be administered as designed: a reply to David Nunan. In T F McNamara (ed.) Language testing colloquium. Selected papers from a Colloquium held at the Horwood Language Centre, University of Melbourne, 24-25 August 1987. Australian Review of Applied Linguistics 11,2: 66-72.

Griffin P E, R J Adams, L Martin and B Tomlinson (1988). An algorithmic approach to prescriptive assessment in English as a Second Language. Language Testing 5,1: 1-18.

Hamp-Lyons L (1989). Applying the partial credit model of Rash analysis: language testing and accountability. Language Testing 6,1: 109-118.

Henning G (1984). Advantage of latent trait measurement in language testing. Language Testing 1,2: 123-133.

Henning G (1987). A guide to language testing: development, evaluation, research. Cambridge, M A: Newbury House.

Henning G (1988). The influence of test and sample dimensionality on latent trait person ability and item difficulty calibrations. Language Testing 5,1: 8-99.

Henning G (1989). Meanings and implications of the principle of local independence. Language Testing 6,1: 95-108.

Henning G and F Davidson (1987). Scalar analysis of composition ratings. In K M Bailey, T L Dale and R T Clifford (eds) Language testing research. Selected papers from the 1986 colloquium. Monterey, CA: Defense Language Institute, 24-38.

Henning G, T Hudson and J Turner (1985). Item response theory and the assumption of unidimensionality for language tests. Language Testing 2,2: 141-154.

Hughes A (1989). Testing for language teachers. Cambridge: Cambridge University Press.

McNamara T F (1987). Assessing the language proficiency of health professionals. Recommendations for the reform of the Occupational English Test. Melbourne: University of Melbourne, Department of Russian and Language Studies.

McNamara T F (1988a). The development of an English as a Second Language speaking test for health professionals. Parkville, Victoria: University of Melbourne, Department of Russian and Language Studies.

McNamara T F (ed.) (1988b). Language testing colloquium. Selected papers from a Colloquium at the Horwood Language Centre, University of Melbourne, 24-25 August 1987. Australian Review of Applied Linguistics 11,2.

McNamara T F (1989a). The development of an English as a Second Language test of writing skills for health professionals. Parkville, Victoria: University of Melbourne, Department of Russian and Language Studies.

McNamara T F (1989b). ESP testing: general and particular. In C N Candlin and T F McNamara (eds) Language, learning and community. Sydney, NSW: National Centre for English Language Teaching and Research, Macquarie University, 125-142.

McNamara T F (1990). Item Response Theory and the validation of an ESP test for health professionals. Paper presented at the Language Testing Research Colloquium, San Francisco, March 2-5.

McNamara T F (in preparation). Assessing the second language proficiency of health professionals. Ph D thesis, University of Melbourne.

Nunan D (1987a). Developing discourse comprehension: theory and practice. Singapore: SEAMEO Regional Langauge Centre.

Nunan D (1987b). Methodological issues in research. In D Nunan (ed.) Applying second language acquisition research. Adelaide: National Curriculum Resource Centre, 143-171.

Nunan D (1988). Commentary on the Griffin paper. In T F McNamara (ed.) Language testing colloquium. Selected papers from a Colloquium held at the Horwood Language Centre, University of Melbourne, 24-25 August 1987. Australian Review of Applied Linguistics 11,2: 54-65.

Oller J W (ed.) (1983). Issues in language testing research. Rowley, M A: Newbury House.

Pollitt A (in press). Diagnostic assessment through item banking. In N Entwhistle (ed.) Handbook of educational ideas and practices. London: Croom Helm.

Pollitt A and C Hutchinson (1987). Calibrated graded assessments: Rasch partial credit analysis of performance in writing. Language Testing 4,1: 72-92.

Skehan P (1989). Language testing part II. Language Teaching 22,1: 1-13.

Spolsky B (1988). Test review: P E Griffin et al. (1986), Proficiency in English as a second language. (1) The development of an interview test for adult migrants. (2) The administration and creation of a test. (3) An interview test of English as a second language. Language Testing 5,1: 120-124.

Spurling S (1987a). Questioning the use of the Bejar method to determine unidimensionality. Language Testing 4,1: 93-95.

Spurling S (1987b). The Bejar Method with an example: a comment on Henning's 'Response to Spurling'. Language Testing 4,2: 221-223.

Widdowson H G (1983). Learning purpose and language use. Oxford: OUP.

Wright B D and G N Masters (1982). Rating scale analysis. Chicago: MESA Press.

Wright B D, R T Congdon and M Rossner (1987). MSTEPS. A Rasch programm for ordered response categories. Chicago, IL: Department of Education, University of Chicago.

THE INTERNATIONAL ENGLISH LANGUAGE TESTING SYSTEM IELTS: ITS NATURE AND DEVELOPMENT

D E Ingram

INTRODUCTION

The international English Language Testing System was formally released for use in November 1989 and is now available in Britain, throughout Australia, in all British Council offices around the world, in the Australian Education Centres and IDP Offices being established in many Asian and Pacific countries, and in other centres where trained administrators are available. The test is the result of a two-year project jointly conducted by Australia and Britain with additional input from Canada and from some individuals elsewhere. The project was funded and directed by the International Development Program of Australian Universities and Colleges, the British Council, and the University of Cambridge Local Examinations Syndicate. The project was led by a team based at the University of Lancaster under the direction of Dr Charles Alderson with the day-to-day activities conducted by Caroline Clapham, supported by a large and varying number of item-writers and reference persons in Britain, Australia, and Canada. A Steering Committee with membership from both countries and including the Project Team oversighted the project in Britain while, in Australia, in addition to teams of item-writers, there was a small Working Party to coordinate Australia's contribution. Australia's input to the Project Team was provided by locating the present writer in Lancaster for thirteenth months in 1987-8 and Dr Patrick Griffin for approximately two months in 1988 while major contributions to the project have continued to be made from Australia throughout the development phase and beyond. The basic reason for the two countries' collaborating to produce the test is that such collaboration shares the inevitably large development cost, draws on a larger and more diverse group of item-writers, and, more importantly, increases considerably the worldwide network of test administration centres available to candidates.

The purpose of the project has been to provide a test suitable for assessing the English proficiency (both general and special purpose) of the large and growing number of international students wishing to study, train or learn English in Australia, Britain, and other English-speaking countries. The test seeks, first, to establish whether such students have sufficient English to undertake training and academic programmes without their study or training being unduly inhibited by their English skills; second, to provide a basis on which to estimate the nature and length of any presessional English course required; third, to provide Australian ELICOS centres (ie, public or private schools providing English Language Intensive Courses for Overseas Students) with a measurement of general proficiency that will indicate the likely level and course needs of students wishing to learn English; and, fourth, when the students exit from an English course, to provide them with an internationally used and recognized statement of English proficiency.

II FORM OF THE IELTS

The form that the IELTS takes has been determined by the purposes it has to serve, evaluations of earlier tests, needs analyses of end-users of the test's results, and significant changes in applied linguistics through the 1980's (especially growing scepticism about the practicality of Munby-type needs analysis (Munby 1978) and ESP theory).

The purposes the IELTS serves require that it provides a valid and reliable measure of a person's practical proficiency both for general and for academic purposes. Because of the large number of students involved, the test has to be readily and rapidly administered and marked, preferably by clerical rather than more expensive professional staff, and able to be administered on demand, worldwide, en masse (except for the Speaking test), and often in remote localities by persons with little training, little professional supervision, and no access to sophisticated equipment such as language laboratories. In addition, once the test has been taken and scored, its results must be readily interpretable even by non-professionals such as institutions' admission officers.

The large number of candidates to whom the test has to be administered also implies a diversity of academic or training fields within which candidates' English proficiency should be measured. While face validity would suggest that the test should have modules relevant to the specific academic or training fields candidates wish to enter, the sheer diversity of those fields (one count, for example, indicated 34 different fields within Engineering) makes a hard version of specific purpose testing impractical while, in any case, there has been growing scepticism through the 1980's about the validity of hard versions of ESP theory and whether the registers of different academic or vocational fields are sufficiently different to warrant different tests in any but very broad domains (see Alderson and Urquhart 1985).

In addition to the constraints imposed by the purposes the test serves, its form was influenced by the pre-development studies made in the course of various evaluations of the British Council's English Language Testing Service (ELTS) (see Clapham 1987, Criper and Davies 1986, ELTS 1987c) and of the needs of end-users of the test's results. These had emphasized the need to limit the total length of the test to less than three hours, the need to provide measures of both general proficiency and ESP/EAP not least in order to ensure face validity, the contrasting need to limit the number of modules in order to simplify administration, the need to provide a profile of results for the four macroskills but to improve the reliability of the Speaking and Writing tests, the desirability of maintaining the use of nine-point band scales and of ensuring that the results (expressed in band scale levels) were readily interpretable, the importance of providing clear and precise specifications to facilitate test replication, and the need for the test to be able to be administered by relatively untrained persons. Changes in testing itself in the last two decades have also played down the value of multiple choice item types and emphasized the need to use types that replicate as far as possible real language use. Considerable thought was given to whether or not the test should include a sub-test specifically focusing on grammar and lexis and, indeed, early versions of the test included a Grammar sub-test which was dropped only after the initial trials and statistical analyses showed that it provided little additional information on a candidate's proficiency.

Because the test is used to assess candidates of a wide range of language proficiency, it was necessary to structure the test to provide measures of proficiency, as far as possible, through the band scale range. This was achieved by having the General component (Listening and Speaking) and the General Training sub-test focus especially around Band level 4 while the EAP/ESP component focuses around Band level 6 without, in both cases, excluding the possibility of information being provided about candidates above and below these levels. In addition, within each sub-test, different phases of the sub-test are focused around different band levels in order to provide graduation within the sub-test as well as between the General and Modular components.

As already indicated, the purposes the IELTS has to serve necessitate that it assess both general proficiency and ESP/EAP proficiency. Consequently, the test has two components. A General Component which has sub-tests in Listening and Speaking and an ESP/EAP component (loosely known as the "Modular" component) which seeks to assess proficiency in Reading and Writing in English for Academic Purposes in one of three broad academic fields (Arts and Social Sciences, Physical Sciences and Technology, and Life and Medical Sciences) and in general proficiency in Reading and Writing in the General Training sub-test. Hence, the overall structure of the IELTS is as appears in Figure 1.

Figure 1: Overall Structure of the IELTS

International English Language Testing System

General Comp.	Band Focus	Modular Comp.	Band Focus
Speaking	4 (3-6)	A Phys. Scs. & Tech	6 (5-7)
Listening	4 (3-6)	B Life & Med. Scs.	6 (5-7)
		C Arts & Soc. Scs.	6 (5-7)
		General Training	4 (3-6)

The nature and purpose of the modular components raise some difficult issues for the form of the test. For reasons already indicated, the three ESP/EAP modules are less specific than, for instance, the former ELTS test and favour the features and academic skills of the broader rather than more specific discipline areas. In addition to the reasons already stated for this, the less specific nature of the Ms and their greater focus on EAP rather than ESP make them more compatible with their use with both undergraduates and graduates. Since some graduates (eg an Engineer taking an MBA) and many undergraduates are often, though not always, just entering a field while many graduates are continuing in their field, there seems at first sight to be some illogicality in using the same ESP-EAP test with both groups but it does seem reasonable to assess whether, whatever their background, applicants to study in a field have both general proficiency and the ability to apply their language to cope with the broad range of academic tasks and register features of the broad discipline area to be entered. However, the M's were considered inappropriate for persons at lower levels of academic development (upper Secondary School), for persons entering vocational training such as apprenticeships, or for persons participating in on-the-job attachments. For these, the emphasis is on general proficiency and consequently they take the G component together with the General Training (GT) module which, like the M's, assesses reading and writing skills but in general and training contexts rather than in discipline-related academic contexts. Persons undertaking training or attachments that make more significant academic demands would take the relevant M rath:r than the GT (eg, persons entering a diploma-level TAFE course, persons going on attachment to a scientific laboratory, or subject-specialist teachers going for practical training, work experience or on exchange in a school). To prevent candidates' taking the General Training module rather than one of the other modules in the belief that it is easier to score higher on the more general test, an arbitrary ceiling of Band 6 has been imposed on the General Training module. The logic, practical value and validity of this decision have yet to be fully tested and will, undoubtedly, be an issue for future consideration by the International Editing committee.

The effect of the pattern of sub-tests just outlined is to enable the IELTS to provide a comprehensive measure of general proficiency in all four macroskillls using the two General component sub-tests in Listening and Speaking and the General Training sub-test in Reading and Writing. For Australian purposes, this makes the test relevant to ELICOS needs where, for persons entering or exiting from a general English course, a comprehensive test of general proficiency is needed. The availability of the EAP/ESP sub-tests in Reading and Writing means that candidates at higher proficiency levels in all four macroskills can also have their proficiency comprehensively assessed though with Reading and Writing being assessed in broad ESP/EAP contexts. It is regrettable that the decision to limit maximum General Training scores to Band 6 prevents persons with high general proficiency in Reading and Writing but no ESP or EAP development from having their skills fully assessed.

III SUB-TESTS AND SPECIFICATIONS:

The current form of the IELTS as released in November 1989 is the result of more than two years' work but the permanent administrative structures allow for considerable on-going development and review activity and, in any case, the test specifications give considerable freedom to item-writers to vary actual item types within clearly stated guidelines. No decision concerning the form and detailed specifications of the test has been taken lightly but, nevertheless, the need to adhere strictly to the original release date of October-November 1989 inevitably meant that some issues will be subject to further trial and investigation and this, together with the continual review process, will mean that the test is not static (and hence rapidly dated) but is in a state of constant evolution.

Detailed specifications have been prepared for all sub-tests so as to facilitate test replication and so that item-writers can, within defined limits, be allowed some flexibility in devising parallel forms of the test. This flexibility is desirable for reasons of test security, to try to ensure that it is proficiency that is being measured and fostered through the washback effect rather than merely the ability to perform specified item types, and to encourage the sort of innovation on the part of item-writers that might lead to progressive improvement in what we would claim is already a valuable test.

The specifications take broadly similar forms subject to the necessary variations arising from the different macroskills. The **Introduction** to all the Specifications outlines the purpose of the IELTS, the nature of the particular test, the target population, the focus of the test (especially in terms of Band Scale levels, tasks, materials, registers and discourse styles), and cultural appropriacy. The next section describes the test structure in detail. In the Speaking test, the structure is described in terms of the length of the interview, the five phases with the purpose, tasks and stimuli, and skills and functions for each. In the Listening test, the structure is described in terms of the time for the test, the stages and, for each stage, the text types, style, utterance rates, item types, skills and functions, contextual features (such as ambient noise, register shifts, number of speakers, accents, and situations), and possible topics. The Ms indicate test focus in terms of band levels, academic tasks, source and audience, the proficiency level, length and structure of texts, and the test tasks that may be included. All Specifications are accompanied by the appropriate Band Scale or, in the case of Writing, Band Scales and by guidelines for non-sexist language.

The Specifications for the International English Language Testing System describe the general nature and purpose of the whole test battery thus:

> "The IELTS is a language test battery designed to assess the proficiency of candidates whose first language is not English and who are applying to undertake study or training through the medium of English. It is primarily intended to *select* candidates who meet specified proficiency requirements for their designated programmes. Its secondary purpose is to be a semi-diagnostic test designed to reveal broad areas in which problems with English language use exist, but not to identify the detailed nature of those problems.
>
> "The test battery consists of a General and a Modular section. The General section contains tests of general language proficiency in the areas of listening and speaking; the Modular section consists of tests of reading and writing for academic purposes."

During the investigations that were conducted prior to the commencement of the development project, informants asked for the Listening sub-test to be included in the modular component because of the importance of students' being able to listen to a lecture, take notes, and carry out a writing task. However, the need to minimize administration time and to be able to administer the test *on* *masse* often without access to a language laboratory made this proposal impractical even for the four broad fields catered for in the modular component of the test. Consequently, Listening is part of the general component and has two stages, the first relating to social situations and the second to course-related situations. A variety of item types may be used including information transfer (such as form-filling, completing a diagram, following routes on a map), more traditional true-false and multiple-choice types, and open-ended (but exhaustively specifiable) answers to questions. Speakers are to be "intelligible native speakers" from one or more of the three countries involved in the project (Australia, Britain and Canada). Discourse styles differ through the test and include conversation, monologue, and formal and informal lectures. Utterance rates are graduated through the test from the lower to middle range of native speaker rates and the contextual features listed ensure that candidates are required to cope with varied accents, different utterance rates, varied but relevant situations, and register shifts.

The Speaking sub-test is discussed in detail elsewhere (Ingram 1990). In brief, it is a direct test of oral proficiency in which the specifications and test outline seek to exercise more control over the interviewer's options than in more traditional approaches using, for example, the ASLPR or FSI Scales. The interview lasts eleven to fifteen minutes, is in five phases, and includes activities that progressively extend the candidate and give him or her the opportunity to lead the discussion. After a short introductory phase in which the interviewer elicits basic personal information, Phase Two gives candidates the opportunity to provide more extended speech about some familiar aspect of their own culture or some familiar topic of general interest. Phase Three uses information gap tasks to have the candidate elicit information and, perhaps, solve a problem. Phase Four draws on a short curriculum vitae filled out by candidates before the interview in order to have them speculate about their future, express attitudes and intentions, and discuss in some detail their field of study and future plans. There is a very short concluding phase entailing little more than the exchange of good wishes and farewell. Assessment is by matching observed language behaviour against a band scale containing nine brief performance descriptions from 1 (Non-Speaker) to 9 (Expert Speaker). Interviewers are native speakers, trained ESL teachers who have undergone short formal training in administering the Speaking test with an additional requirement to work through an interviewer training package at regular intervals. All interviews are audio-recorded with a 10% ample being returned to Australia or Britain for monitoring of interview quality and moderation of assessments assigned. This sub-test is of particular interest from a test design point of view since it draws on the developments in "direct", interview-based assessment of the last two decades but seeks to control the interview to maximize validity and reliability for large-scale administration often in remote locations using minimally trained interviewers. Of particular importance is the attempt made in the interview to surrender initiative from the interviewer to the candidate (especially in Phase 3) because of the importance, in English-speaking academic environments, of students' being willing to ask questions and seek information for themselves.

Three of the modular tests assess reading and writing in ESP-EAP contexts while the General Training module (to be discussed subsequently) assesses them in general and training contexts. It was noted earlier that these tests each have to be appropriate for candidates in a wide range of disciplines and for both those entering and those continuing their field. Consequently, though reading or stimulus materials are chosen from within the broad discipline areas of the target population of the module, neutral rather than highly discipline-specific texts are to be used, which excludes such materials as is found in textbooks so that item-writers are required to choose "(scientific) magazines, books, academic papers and well-written newspaper articles (in a relevant field) written by scientists for the informed lay person and for scientists in other fields". Reading and writing are integrated so that, as is integral to academic writing, at least one of the writing tasks draws on material in one or more of the reading texts. The difficulty level of the texts is to be within the target proficiency range (Bands 5 to 7) except that, where the Writing test draws on reading materials for input to the writing task, the difficulty level should not exceed Band 5 so as to minimize the chance of reading interfering with the candidate's ability to demonstrate writing proficiency. The reading tasks focus on tasks relevant to academic reading and include, amongst others, identifying structure, content and procedures, following instructions, finding main ideas, identifying underlying themes, evaluating and challenging evidence, reaching a conclusion, and drawing logical inferences. Though the reading test items have to be clerically markable, they make use of a variety of item types including, *inter alia*, cloze-type procedures, summary completion, table completion, heading insertion, multiple choice, and exhaustively specifiable open-ended questions.

The writing tasks are as realistic as possible and require the sort of activity that candidates will have to do on entering and pursuing their courses, including, amongst others, organizing and presenting data, describing an object or event, explaining how something works, presenting and justifying an opinion or hypothesis, comparing and contrasting evidence, arguing a case, and evaluating and challenging evidence. The assessment procedure for the Writing test is the most complicated in the test battery and uses three different sub-scales for each of the two tasks in the Writing test. These have then to be converted to a single band scale score. Whether this more analytic approach to assessing writing proficiency is more global band scale or whether the increased complexity and time demands of the scoring procedure mitigate any possible benefit from the analytic approach has yet to be convincingly demonstrated by research and this will undoubtedly be one of the urgent matters to be reconsidered by the

International Editing Committee now that the test has been formally released. The approach to writing assessment used in the IELTS is to be discussed in a workshop presentation elsewhere in the 1990 RELC Seminar.

The General Training test focuses on Bands 3 to 6 (rising through the test) with, as already noted, a ceiling of Band 6 on scores that can be assigned. The tasks are required to focus on those skills and functions relevant to survival in English speaking countries and in training programmes including, amongst others, following and responding to instructions, identifying content and main ideas, retrieving general factual information, and identifying the underlying theme or concept. Texts should not be specific to a particular field, may be journalistic, and should include types relevant to training contexts and survival in English speaking countries (eg, notices, posters, straightforward forms and documents, institutional handbooks, and short newspaper articles). Some of the possible item types in the reading test may include inserting headings. information transfer, multiple choice, short-answer questions with exhaustively specifiable answers, and summary completion. In the Writing test, the skills and functions that may be tested include presenting data, giving instructions, expressing needs and wants, expressing an opinion, providing general factual information, and engaging in personal correspondence in a variety of topics relevant to training and survival in an English speaking country.

Some examples of item types have been given in discussing the form of the sub-tests. Their choice is constrained by several factors that are not always compatible, including economy of administration and scoring and the need to provide a realistic measure of the candidate's practical proficiency. To reduce costs, the tests are designed to be clerically, and hence essentially, objectively, markable except for Speaking and Writing which are assessed directly against Band Scales by ESL teachers trained and accredited to assess. Some use of multiple-choice questions remains though all items are contextualized and item-writers are required at all times to consider the realism of the task the candidate is being asked to undertake. Item-writers are also required at all times to relate texts, items and anticipated candidate responses to the Band Scales so as to facilitate interpretation, to provide the required band scale focus, and to ensure an appropriate gradation through the test.

The proficiency level of the test and its component parts is a critical issue determining its overall usefulness. It was important, in order for the test to meet the variety of needs it was intended to cover, that it have a wide proficiency range and provide a measure of proficiency in all four macroskills. The General component and General Training consequently focus around Band 4 and the three academic sub-tests focus around Band 6 with the band scale spread shown in Figure 1 (without excluding the possibility in all sub-tests of information being provided on candidates above or below these levels) but, in addition and as already indicated, the sections within each sub-test are graduated to enable each to cater for a wider range of candidates. In other words, though the IELTS is designed to select candidates for academic courses and training programmes, it does not just adopt a threshold approach but is designed to make a statement about candidates' proficiencies whatever their level might be.

It was indicated earlier that it was a requirement on the project Team to keep the total length of the IELTS below three hours. The actual length of the current form of the test is a maximum of 145 minutes made up of 30 minutes for the Listening test, the Speaking test is 11 to 15 minutes, and the Ms are 100 minutes with 55 minutes for Reading and 45 for Writing.

IV THE INTERNATIONAL NATURE OF THE TEST:

The new test, as the name indicates, is the result of an international project involving three English-speaking countries with their own distinctive dialects and cultures. The test has to be seen to be an international test and to be appropriate to all candidates whatever their country of origin and target country. This has significantly affected the operation of the development project and the content of the test and is a continual factor in determining the permanent management structures and technical operations.

In development and administration, the IELTS has had to recognize the respective roles and contributions of the participating countries (especially Australia and Britain, the principal participants, and now the owners of the test), to recognize and utilize the expertise in those countries, and to be clearly the result of the equal-status collaboration of the participating countries. It has been necessary, therefore, to ensure major contributions to all phases of the project from both Australia and Britain even though the development project was based in Lancaster. Hence, the present writer was based in Lancaster for thirteen months and Patrick Griffin for six weeks, an Australian Working Party provided additional support, and major contributions continued from Australia throughout 1989 after the present writer returned to Australia from Lancaster. The permanent management, development and administration arrangements for the test to be referred to below are also designed to ensure equal ownership and involvement by both Australia and Britain. Indeed, in the unlikely event that either country were to take action that would relegate the role of the other to a more subordinate position in management, development, monitoring, or supporting research, the international arrangements would probably cease and, in Australia's case, a new Australian test acceptable to Australian institutions and the large ELICOS market would have to be devised. There is no reason to anticipate such an unsatisfactory event occurring since the project has proceeded harmoniously and effectively and the point is made here solely to stress the importance placed on the international nature of the test, its development, and its on-going management. A number of factors arising from the internationalism of the test are worthy of note.

First, participating countries have had to adjust their requirements for test content to accommodate each other's needs. Thus, for instance, the non-academic module in the former ELTS test has been changed to meet Australian ELICOS needs so that the IELTS General Training module includes a lower proficiency range, and focuses mainly around general proficiency in reading and writing.

Second, care is taken to ensure that place names or anything else identified with one country are balanced by features identified with the other and unnecessary references to one or the other country are avoided so as to ensure that nothing is done that could lead the test to be associated with any one of the participating countries or to bias the test towards candidates going or going back to any of the countries.

Third, in language and culture, the test focuses on international English rather than on British, Australian or Canadian English. On the cultural level, it is reasonable to expect that candidates learning English will learn the general culture that underlies the language and constitutes the meaning system. However, an international test must avoid country-specific assumptions based on just one of the cultures involved. To do this, however, requires careful editing by testers from all participating countries since it is often impossible for a person, however expert but locked in his or her own culture, even to recognize when knowledge of certain aspects of his or her culture is assumed. On one occasion during the writing of an early item in a draft Listening test, for example, the present writer, coming from Australia where mail is delivered once a day, Monday to Friday, interpreted a question and marked a multiple choice answer quite differently from what was intended by the British writer who assumed that mail was normally delivered once or twice a day, six days a week. In another question in another sub-test, a person familiar with Australian culture would have marked a different multiple-choice answer because of assumptions arising from the fact that, in Australia, large numbers of sheep are trucked on road trains and so a large flock could be "driven" from one place to another whereas, if they had been walked there (which the British test writer intended), the verb used would have been "to drove".

Fourth, the Specifications for all the sub-tests include a section on cultural appropriacy which emphasizes the need for the test to be equally appropriate for students going to any of the participating countries and the need to avoid country-specific cultural knowledge and lexical and other items identified with any one variety of English. The section on cultural appropriacy also emphasizes the need to avoid topics or materials that may offend on religious, political or cultural grounds and to observe international guidelines for non-sexist language.

V MANAGEMENT AND ADMINISTRATION

As already indicated, the IELTS is available worldwide from British Council offices, Australian Education Centres, offices of the International Development Program of Australian Universities and Colleges (IDP), and from other places administered by the British Council or IDP where trained administrators, interviewers and markers are available. Where both Australian and British offices are available, the centres are required to work cooperatively rather than in competition. Both countries are responsible for the training of administrators, interviewers and markers working to an agreed training schedule and both countries will take responsibility for the monitoring and moderation of interviewing and the rating of Speaking and Writing proficiencies. Both countries will also cooperate in other aspects of IELTS administration, management, technical development, and related research and the supporting structures have been established to ensure that this occurs.

The senior management body is the International Management Committee with equal representation from Britain and Australia. The senior technical body is the International Editing Committee which consists of the two (ie, Australian and British) Chief Examiners and a chairperson. The chairs of the IMC and IEC are to alternate between Britain and Australia, with one being Australian and the other British. In Britain, the test is owned by the British Council which has contracted its management and the technical aspects of the on-going test development to the University of Cambridge Local Examinations Syndicate while, in Australia, a consortium of institutions, called "IELTS Australia", has been established with IDP holding half the shares. The item-writing is carried out equally in both countries by teams of item-writers engaged for the purpose with items being submitted to the national Chief Examiner and subsequently to the International Editing Committee before formal trialling takes place. At least one parallel form of the test will be developed each year though it is strongly to be hoped that this number will rapidly increase in order to ensure test security.

VI CONCLUSION

The International English Language Testing System is a test of some considerable interest for several reasons. First, it will rapidly become (if it is not already) the principal test of English proficiency to be taken by students or trainees going to Australia and Britain to study, train or learn English. Second, this is probably the first time that a major test has been developed and maintained in an international project of this sort, in which, in addition to technical cooperation, the project has sought and continues to seek to draw equally on the testing expertise available in the participating countries, to foster that expertise on a wide scale by the deliberate involvement of applied linguists across each nation, and to develop a test compatible with the needs, dialects and cultures of the participating countries. Third, certain features of the test itself are of interest, not least the structured controls on the Speaking test and the attempt to give candidates the opportunity to take initiative during the interview. Fourth, there has been a deliberate attempt throughout the development process to consider the washback effect of the test on English language teaching in other countries and to adopt test techniques that are more likely to have a favourable influence on the teaching of English. Finally, the sheer magnitude of the project, the large number of candidates that will be taking the test, and the need for much on-going test monitoring and regeneration will provide a considerable stimulus to the development of language testing as a skilled activity in both countries. It is, for instance, not coincidental that the recently established Languages Institute of Australia (a nationally funded centre for research and information in applied linguistics in Australia) includes within it two language testing units established in two of the institutions most involved in Australia's contribution to the development of the International English Language Testing System.

REFERENCES

Alderson, J Charles and Caroline Clapham, 1987, *"The Revisiion of the ELTS Test". Paper to the English Language Education Forum, University of Lancaster, 3 November, 1987.*

Alderson, J Charles and A H Urquhart, 1985. *"This Test is Unfair: I'm not an Economist". In Hauptman et al 1985.*

Burke, Ed, 1987. *"Final Report on the Evaluation of Overseas Standardized English Language Proficiency Tests: Implications for English Language Proficiency Testing in Australia". Report to the Overseas Students office, Commonwealth Department of Education, Canberra, Australia, 30 April. 1987.*

Clapham, Caroline, 1987. *"The Rationale for the Structure of the Revised ELTS". Mimeograph.*

Criper, Clive and Alan Davies, 1986. *"Edinburgh ELTS Validation Project: Final Report". Mimeograph. Subsequently published by Cambridge University Press, 1988.*

English Language Testing Service (ELTS), 1987. *An Introduction to ELTS, London: British Council.*

English Language Testing Service (IELTS), 1987a. *Specimen Materials Booklet, London: British Council.*

English Language Testing Service (ELTS), 1987b. *Non-Academic Training Module: User Handbook. London: British Council.*

English Language Testing Service, 1987c. *"Acceptability of ELTS in UK Universities and Polytechnics". London: British Council.*

Hauptman, P C, R LeBlanc and M B Wesche (eds), 1985. *Second Languege Performance Testing, Ottawa: University of Ottawa Press.*

Ingram, D E. 1990. *"The International English Language Testing System: The Speaking Test", workshop presentation to the 1990 RELC Seminar, Singapore, 9-13 April 1990.*

Ingram, D E. *In preparation. Direct Proficiency Assessment.*

Ingram, D E and Caroline Clapham, 1988. *"ELTS Revision Project: A New International Test of English Proficiency for Overseas Students". Paper to the 16th FIPLV World Congress on Language Learning/7th Biennial AFMLTS National Languages Conference, "Learning Languages is Learning to Live Together", Australian National University, Canberra, 4-8 January, 1988.*

Lo Bianco, Joseph, 1987. *National Policy on Languages, Canberra: Australian* Government Publishing Service.

Munby, J L, 1978. *"Communicative Syllabus Design", Cambridge University Press.*

A COMPARATIVE ANALYSIS OF SIMULATED AND DIRECT ORAL PROFICIENCY INTERVIEWS

Charles W. Stansfield

This article introduces the reader to the simulated oral proficiency interview and discusses the research that has been conducted on it to date. Subsequently, it compares this type of test with a face-to-face interview in respect to reliability, validity, and practicality. Finally, it offers some reasons why the simulated oral proficiency interview is as good a measure of oral language proficiency as the face-to-face interview and describes the situations in which it may actually be preferable to the face-to-face format.

INTRODUCTION

The simulated oral proficiency interview (SOPI) is a type of semi-direct speaking test that models, as closely as is practical, the format of the oral proficiency interview (OPI). The OPI is used by US Government agencies belonging to the Interagency Language Roundtable (ILR) and by the American Council for the Teaching of Foreign Language (ACTFL) to assess general speaking proficiency in a second language. The OPI, and the scale on which it is scored, is the precursor of the Australian Second Language Proficiency Rating (ASLPR).

The measure I have called a SOPI (Stansfield, 1989) is a tape-recorded test consisting of six parts. It begins with simple personal background questions posed on the tape in a simulated initial encounter with a native speaker of the target language. During a brief pause, the examinee records a short answer to each question. Part one is analogous to the "warm-up" phase of the OPI. The remaining five parts are designed to elicit language that is similar to that which would be elicited during the level check and probe phases of the OPI. Parts two, three, and four employ pictures in a test booklet to check for the examinee's ability to perform the various functions that characterize the Intermediate and Advanced levels of the ACTFL proficiency guidelines, or levels one and two of the ILR skill level descriptions. Thus, the examinee is asked to give directions to someone using a map, to describe a particular place based on drawing, and to narrate a sequence of events in the present, past, and future using drawings in the test booklet as a guide. Parts five and six of the SOPI require examinees to tailor their discourse strategies to selected topics and real-life situations. These parts assess the examinee's ability to handle the functions and content that characterize the Advanced and Superior levels of the ACTFL guidelines, or levels two through four of the ILR skill level descriptions. Like the OPI, the SOPI can end with a wind-down. This is usually one or more easy questions designed to put the examinee at ease and to facilitate the ending of the examination in as natural a manner as possible.

After the test is completed, the tape is scored by a trained rater using the ACTFL/ILR scale. The score an examinee earns may range from the Novice level to High Superior (See Figure 1). The Novice level is equivalent to level 0 or 0+ on the ILR scale, while High Superior is equivalent to a rating of between 3+ and 5 on the ILR scale.

RESEARCH AND DEVELOPMENT INVOLVING THE SOPI

In five studies involving different test development teams and different languages, the SOPI has shown itself to be a valid and reliable surrogate of the OPI. Clark and Li (1986) developed the first SOPI, although they did not label it as such, in an effort to improve on the Recorded Oral Proficiency Interview, or ROPE test, which was a semi-direct version of the OPI containing instructions and questions entirely in the target language (Lowe and Clifford, 1988). Clark and Li developed four forms of a ROPE-like test of Chinese, with instructions and scenarios in English, and then administered the four forms and an OPI to 32 students of Chinese at two universities. Each test was scored by two raters and the scores on the two types of test were statistically compared. The results showed the correlation between the SOPI and the OPI to be .93.

Shortly after arriving at the Center for Applied Linguistics (CAL) in 1986, I read Clark's report on this project and realized that these favorable results merited replication by other researchers in situations involving other test developers and learners of other languages. As a result, I applied for a grant from the US Department of Education to develop similar tests in four other languages. Fortunately, the grant was funded, and in August 1987 I began the development of a similar semi-direct interview test of Portuguese, called the Portuguese Speaking Test (Stansfield, et al., 1990).

Three forms of this test and an OPI were administered to 30 adult learners of Portuguese at four institutions. Each test was also scored by two raters. In this study a correlation of .93 between the two types of test was also found. In addition, the SOPI showed itself to be slightly more reliable than the OPI, and raters reported that the SOPI was easier to rate, since the format of the test did not vary with each examinee.

During 1988 and 1989, I directed the development of tests in Hebrew, Hausa, and Indonesian. The Hebrew SOPI, or Hebrew Speaking Test (HeST) as we call it, was developed in close collaboration with Elana Shohamy and her associates at the University of Tel Aviv (Shohamy et al., 1989). In order to accommodate the different settings where the language is studied and used, two forms of the test were developed for use in Hebrew language schools for immigrants to Israel, and two forms were developed for use in North America. The first two forms were administered to 20 foreign students at the University of Tel Aviv and the other two forms were administered to 10 students at Brandeis University and 10 students at the University of Massachusetts at Amherst. Each group also received an OPI. The correlation between the OPI and this SOPI for the Israeli version was .89, while the correlation for the U S version was .94. Parallel-form and interrater reliability were also very high. The average interrater reliability was .94 and parallel form reliability was .95. When examinees' responses on different forms were scored by different raters, the reliability was .92.

Recently, Dorry Kenyon (my associate at CAL) and I reported on the development and validation of SOPIs in Indonesian and Hausa (Stansfield and Kenyon, 1989). The development of the Indonesian Speaking Test (IST) posed special problems. Indonesian is one of those languages where the context of the speech situation seems to be especially important. Because of this, we strived to contextualize the test items to an even greater degree than had been done for other languages. In order to do this, we specified the age, sex, and position or relationship of the supposed interlocutor for the examinee. During trialing, we noticed that examinees tended to assign a name to the person they were speaking with. As a result, we gave each interlocutor, as appropriate, a name on the operational forms. To validate the test, 16 adult learners of Indonesian were administered two forms of the IST and an OPI. The correlation with the OPI was .95. Reliability was also high, with interrater reliability averaging .97, and parallel-form reliability averaging .93 for the two raters. When different forms and different raters were used, the reliability was also .93.

The development of two forms of the Hausa Speaking Test also posed special problems. Here, it was necessary to develop a version for male examinees and a version for female examinees, because the pronoun "you" carries gender in Hausa as it does in Hebrew. Because no ACTFL or ILR-certified interviewer/raters were available for Hausa, it was not possible to administer an OPI to the 13 subjects who took the Hausa Speaking Test.

However, two speakers of Hausa as a second language who had received familiarization training in English with the ACTFL/ILR scale, subsequently scored the Hausa test tapes on that scale. The raters showed high interrater reliability (.91) in scoring the test and indicated that they believed it elicited an adequate sample of language from which to assign a rating.

COMPARATIVE CHARACTERISTICS OF THE SOPI AND THE OPI

A comparison of the two types of test demonstrates that the SOPI can offer a number of advantages over the OPI with respect to the fundamental psychometric characteristics of reliability, validity and practicality.

Reliability. The SOPI has shown itself to be at least as reliable and sometimes more reliable than the OPI. During the development of the Chinese Speaking Test (Clark and Li, 1986) the OPI showed an interrater reliability of .92, while the four forms of the SOPI showed an interrater reliability of .93. On the Portuguese SOPI that I developed, the interrater reliability for three forms varied from .93 to .98, while the reliability of the OPI was .94. In addition, some raters reported that it was sometimes easier to reach a decision regarding the appropriate score for an examinee who was taking the SOPI than for an examinee who was taking the OPI. This is because the OPI requires that each examinee be given a unique interview, whereas the format and questions on an SOPI are invariant. Under such circumstances, it is often easier to arrive at a decision on the score. The situation is similar to scoring a batch of essays on the same topic versus scoring essays on different topics. The use of identical questions for each examinee facilitates the rater's task. I should be careful to point out that although the rater's task is made easier by the use of identical questions, competent raters are able to apply the scale reliably when different questions are used. Thus, the use of a common test for all examinees does not guarantee an improvement in reliability over the face-to-face interview.

The length of the speech sample may also facilitate a decision on a rating. The OPI typically takes about 20 minutes to administer and produces about 15 minutes of examinee speech. The SOPI takes 45 minutes to administer and produces 20-23 minutes of examinee speech. Thus, there is a greater sample of performance for the rater to consider on the SOPI and this sample may make distinctions in proficiency more salient.

Another advantage is found in the recording of the test for later scoring. In the OPI, the same interviewer typically rates and scores the test. Yet this interviewer may not be the most reliable or accurate rater. In the SOPI, one can have the tape scored by the most reliable rater, even if this rater lives in a different city or region of the country.

Validity. Many factors can affect the validity of a measure of oral proficiency. The consideration of several factors explains why the SOPI may be as valid as the OPI.

The SOPI usually produces a longer sample of examinee speech. When this is the case, the more extensive sample may give it greater content validity.

In an OPI, the validity of the speech sample elicited is in large part determined by the skill of the interviewer. If the interviewer does not adequately challenge the examinee by posing demanding questions, the examinee will not be given a chance to demonstrate his or her language skills. If the interviewer consistently asks questions that are too demanding for the examinee, then the examinee's language skills may appear to be consistently faulty on all tasks, with the result that a lower score may be assigned than is warranted. Similarly, the interviewer may miss opportunities to question the examinee about topics that are of personal interest or within his or her range of awareness. Or, the interviewer and the interviewee may have very little in common. Finally, if the interview is too short, it will not adequately sample the language skills of the interviewee. All of these factors can affect the validity of the OPI.

Although interviewers can vary considerably in their interviewing techniques, the SOPI offers the same quality of interview to each examinee. Parallel forms of the SOPI can be developed with great care over a period of time, so as to ensure that they are comparable in quality and difficulty. The parallel forms developed thus far have shown nearly identical correlations with OPIs administered by highly trained interviewers. Thus, different forms of the SOPI, unlike different interviewers, appear to be equal in validity, even when rated by different raters.

Many second language educators feel that the face-to-face OPI is the most valid test available. Thus, it is appropriate to consider the effects of the SOPI's semi-direct format on its validity as a measure of general oral language proficiency. One point of comparison is the naturalness with which topics are switched during the test. Within the context of the SOPI, the topic changes with each question in Parts II through VI, for a total of approximately 15 transitions, depending on the language of the test. Yet because of the test-like format of a semi-direct measure, the change in topic seems perfectly natural to the examinee. In the OPI, the examiner must change the topic on a number of occasions in order to provide adequate sampling of the content. This switching of topic, if done too abruptly, can seem awkward and disconcerting to the interviewee. This is not the case when the topic is switched naturally, but such natural changes in topic of the conversation can only be brought about a limited number of times (4-8) within the span of a 20 minute conversation. As a result, the SOPI makes possible a greater number of topical transitions, which contribute to greater content sampling on the part of the SOPI.

Another point of comparison between the two test formats is the role play situation. Usually, the OPI includes two role plays. These are usually presented to the interviewee on a situation card, written in English. The interviewee reads the card to the interviewer and then both interlocutors play the roles prescribed on the card. Although somewhat artificial, these situations are incorporated into the interview because they provide useful diagnostic information on the strengths and weaknesses of the interviewee. Yet only two situations are included in the OPI. The SOPI includes five situations in Part VI, thereby providing a greater amount of diagnostic information than the OPI.

Since speaking into a tape recorder is admittedly a less natural situation than talking to someone directly, it is possible that the SOPI format will cause undue stress. However, feedback from examinees has not indicated that this is the case. While most examinees prefer the face-to-face interview, because of the human contact it provides, about a quarter of the examinees either have no preference or actually prefer to speak into a tape recorder. The latter group claim they feel less nervous than when forced to converse face-to-face with an unfamiliar and highly competent speaker of the target language.

One may also examine the test situation itself as a source of unnaturalness. In the OPI the examinee speaks directly to a human being. However, the examinee is fully aware that he or she is being tested, which automatically creates unnatural circumstances. As van Lier (1989) has noted, in the OPI the aim is to have a successful interview, not a successful conversation. Thus, even the OPI is not analogous to a real conversation. The SOPI, on the other hand, would seem even less natural, since it is neither a conversation nor an interview. In short, neither format produces a "natural" or "real-life" conversation.

As mentioned above, the interview usually contains two role plays that are described to the examinee on situation cards printed in English. During this portion of the interview, the exam:nee is fully aware that the examiner is not a waiter, a hotel clerk, a barber, a cab driver, or the next door neighbor. Yet the examinee has to engage in spontaneous acting with the interviewer in order to succeed. The situational portion of the SOPI may be actually more natural than in the OPI, since the examinee is free to imagine that he or she is talking to the people described in the situation prompt.

In the SOPI format, the aim of the interviewee is to perform as well as possible on the test. Unnaturalness seems to be a natural part of the test situation. Tests themselves are unnatural samples of examinee performance. This is a fundamental reason why the validity of test scores is always an important issue. Tests, whether direct, semi-direct, or indirect, are mere indicators of the true underlying ability they claim to measure. Yet tests can be valid measures of this ability, whether they are natural in format or not.

Further examination of the nature of the OPI gives critical clues as to why the SOPI correlates so highly with it, even when the OPI is conducted by experienced, expert interviewers. The explanation probably lies in the limitations of the OPI itself. Since the SOPI does not measure interactive language, and the two tests measure the same construct, then the examinee's skill in verbal interaction must not play a significant role on the OPI. Consideration of the relationship between interviewer and interviewee on the OPI suggests this is indeed the case. The interviewer typically asks all the questions and maintains formal control over the direction of the conversation. The interviewee plays the subservient role, answering questions and responding to prompts initiated by the interviewer with as much information as possible. He or she has little if any opportunity to ask questions, to make requests, exclamations or invitations. Nor does the interviewee have the opportunity to demonstrate sociolinguistic competence in a variety of situations, such as when speaking to member of the opposite sex, older and younger persons, or individuals of higher or lower status. The interviewer is trained to maintain a secondary profile, and to not engage in back-and-forth discussion or exchange with the examinee. Both parties understand that it is the examinee's responsibility to perform. Little true interaction takes place.

The lack of authentic interaction in the OPI prompted van Lier (1989) to state: "Since it is so difficult to attain conversation in the formal context of an OPI and since we have not developed sufficient understanding of what makes conversation successful in order to conduct reliable and valid ratings, it would be easier for all concerned if we could dispense with conversation as the vehicle for evaluation" (p. 501). I do not propose dispensing with the OPI. However, given the lack of true interaction in the OPI, it is not surprising that the SOPI and the OPI correlate so well.

It should be noted that there may be circumstances where interactive skills or pragmatic or sociolinguistic competence need to be measured. In such circumstances, the OPI would appear to be potentially more useful. However, in order to do this one would have to modify the OPI to focus on these abilities. One would also have to modify the scale, so that it would reflect the examinee's interactive ability. Or, perhaps it would be more appropriate to assign a separate rating for interaction.

Perhaps a greater understanding of the two test types can be gleaned from qualitative research into examinees' performance on them. If a content analysis or discourse analysis of examinee speech indicated that either format elicits a wider spectrum of language skills, then additional content validity would accrue to that format. Similarly, if the two test types seem to elicit language that is qualitatively different, then it would be helpful to know this as well. Currently, we have available tapes containing examinee responses under both formats. Elana Shohamy and her associates are currently planning a qualitative study of the Hebrew tapes. We are willing to make the tapes in Chinese, Portuguese, Hausa and Indonesian available to other serious researchers. The results of such studies have the potential to contribute greatly to our understanding of the validity of each type of test.

Practicality. The SOPI offers a number of practical advantages over the OPI. The OPI must be administered by a trained interviewer, whereas any teacher, aide, or language lab technician can administer the SOPI. This may be especially useful in locations where a trained interviewer is not available. In the US, this is often the case in languages that are not commonly taught, which are those for which I have developed SOPI tests thus far.

Another advantage is that the SOPI can be simultaneously administered to a group of examinees by a single administrator, whereas the OPI must be individually administered. Thus, the SOPI is clearly preferable in situations where many examinees need to be tested within a short span of time.

The SOPI is sometimes less costly than the OPI. If a trained interviewer is not available locally, one will have to be brought to the examinees from a distance, which can result in considerable expenditure in terms of the cost of travel and the interviewer's time. The fact that the SOPI makes it possible to administer the test simultaneously to groups obviates the need for sveral interviewers who would interview a number of examinees within a short period of time.

CONCLUSION

An examination of the SOPI research, which has been carried out on different examinees, and on tests of different languages produced by different test development teams, shows that the SOPI correlates so highly with the OPI that is seems safe to say that both measures test the same abilities. The SOPI has also shown itself to be at least as reliable as the OPI, and in some cases more so. Thus, it seems safe to conclude that it is as good as an OPI in many situations. Furthermore, a comparison of the advantages of each has shown that the SOPI can offer certain practical and psychometric advantages over the OPI. Thus, it may be useful to consider the circumstances that should motivate the selection of one format or the other.

Since the tasks on the SOPI are ones that can only be effectively handled by responding in sentences and connected discourse, the SOPI is not appropriate for learners below the level of Intermediate Low on the ACTFL scale or level 1 on the ILR scale, since examinees whose proficiency is below this level use words and memorized phrases, not sentences, to communicate. Similarly, the standardized, semi-direct format of the test does not permit the extensive probing that may be necessary to distinguish between the highest levels of proficiency on the ILR scale, such as levels 4, 4+, and 5.

The purpose of testing may also play a role in the selection of the appropriate format. If the test is to have very important consequences, it may be preferable to administer a SOPI, since it provides control over reliability and validity of the score. Such would seem to be the case when language proficiency will be used to determine whether or not applicants are qualified for employment. Examples of such important uses are the certification of foreign trained medical personnel and the certification of foreign language and bilingual education teachers. (The Texas Education Agency, which is the coordinating agency for public schools in the state of Texas, agrees with me on this point. Recently, it awarded CAL a contract to develop SOPI tests in Spanish and French for teacher certification purposes in Texas).

When conducting research on language gains or language attrition, use of the SOPI would permit one to record the responses of an examinee at different points in time, such as at six months intervals. These responses could then be analyzed in order to determine their complexity. In this way, the SOPI would serve a valid measure of general language competence, while allowing the researcher to completely standardize the test administration. Many other research situations requiring a valid and reliable measure of general oral language proficiency, would also seem to call for the SOPI.

When scores will not be used for important purposes, and a competent interviewer is available, it would seem preferable to administer an OPI. Such is often the case with placement within an instructional program. In such a situation, an error in placement can be easily corrected. Similarly, an OPI administered by a competent interviewer may be preferable for program evaluation purposes because of the qualitative information it can provide and because the score will not have important repercussions for the examinee. Ultimately, the type of test chosen will depend on the purpose for testing, and on practical considerations.

It may appear that I am suggesting that the OPI is not a valid and reliable test. This is not the case. I continue to view the OPI as potentially being the more valid and reliable measure when carefully administered by a skilled interviewer and rated by an accurate rater. I also recognize that the OPI can assess a broader range of examinee abilities that can the SOPI. The central point I have made here is that when quality control is essential, and when it can not be assured for all examinees using the OPI, then the SOPI may be preferable, given the high degree of quality control it offers. When quality control can be assured, or when it is not a major concern, or when assessment at very low and very high ability levels is required, or when practical considerations do not dictate test type, then the OPI may be preferable.

REFERENCES

Clark, J. L. D. (1979). *Direct vs. semi-direct tests of speaking ability. In E. J Briere and F. B. Hinofotis (Eds.), Concepts in Language Testing: Some Recent Studies (pp. 35-49). Washington, DC: Teachers of English to Speakers of Other Languages.*

Clark, J. L. D. and Li, Y. (1986). *Development, Validation, and Dissemination of a Proficiency-Based Test of Speaking Ability in Chinese and an Associated Assessment Model for Other Less Commonly Taught Languages. Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 278 264).*

Clark, J. L. D and Swinton, S. S. (1979). *Exploration of Speaking Proficiency Measures in the TOEFL Context (TOEFL Research Report 4). Princeton, NJ: Educational Testing Service.*

Lowe, P. and Clifford, R. T. (1980). *Developing an Indirect Measure of Overall Oral Proficiency. In, J. R. Frith, Editor, Measuring Spoken Language Proficiency. Washington, DC: Georgetown University Press.*

Shohamy, E., Gordon, C., Kenyon, D. M., and Stansfield, C. W. (1989). *The Development and Validation of a Semi-Direct Test for Assessing Oral proficiency in Hebrew. Bulletin of Hebrew Higher Education, 4(1), pp. 4-9.*

Stansfield, C. W. (1989). *Simulated Oral Proficiency Interviews. ERIC Digest. Washington, DC: ERIC Clearinghouse on Languages and Linguistics.*

Stansfield, C. W and Kenyon, D. M. (1988). *Development of the Portuguese Speaking Test. Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 296 586).*

Stansfield, C. W and Kenyon, D. M. (1989). *Development of the Hausa, Hebrew, and Indonesian Speaking Tests. Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service, Forthcoming).*

Stansfield, C. W, Kenyon, D. M, Paiva, R, Doyle, F., Ulsh, I., and Cowles, M. A. (1990). *The Development and Validation of the Portuguese Speaking Test. Hispania, 73(3), 641-651.*

Van Lier, L. (1989). *Reeling, Writing, Drawling, Stretching, and Fainting in Coils: Oral Proficiency Interviews as Conversation. TESOL Quarterly, 23(3), 489-508.*

Figure 1. SOPI Scale.

NOVICE	The Novice level is characterized by the ability to communicate minimally with learned material. The PST is designed for examinees who exceed this level. Any examinee not achieving the minimum ability to be rated at the Intermediate level will receive this rating.
INTERMEDIATE	The Intermediate level is characterized by the speaker's ability to • create with the language by combining and recombining learned elements, though primarily in a reactive mode. • initiate, minimally sustain, and close in a simple way basic communicative tasks, and • ask and answer questions.
Intermediate Low	Able to handle successfully a limited number of interactive, task-oriented and social situations. Misunderstandings frequently arise, but with repetition, the Intermediate Low speaker can generally be understood by sympathetic interlocutors.
Intermediate Mid	Able to handle successfully a variety of uncomplicated, basic and communicative tasks and social situations. Although misunderstandings still arise, the Intermediate Mid speaker can generally be understood by sympathetic interlocutors
Intermediate High	Able to handle successfully most uncomplicated communicative tasks and social situations. The Intermediate-High speaker can generally be understood even by interlocutors not accustomed to dealing with speakers at this level, but repetition may still be required.
ADVANCED	The Advanced level is characterized by the speaker's ability to • converse in a clearly participatory fashion - initiate, sustain, and bring to closure a wide variety of communicative tasks, including those that require an increased ability to convey meaning with diverse language strategies due to a complication or an unforeseen turn of events; • satisfy the requirements of school and work situations, and • narrate and describe with paragraph-length connected discourse.
Advanced-Plus	In addition to demonstrating those skills characteristic of the Advanced level, the Advanced Plus level speaker is able to handle a broad variety of everyday, school, and work situations. There is emerging evidence of ability to support opinions, explain in detail, and hypothesize. The Advanced-Plus speaker often shows remarkable fluency and ease of speech but under the demands of Superior-level, complex tasks, language may break down or prove inadequate.
SUPERIOR	The Superior level is characterized by the speaker's ability to: • participate effectively and with ease in most formal and informal conversations on practical, social, professional, and abstract topics; and • support opinions and hypothesize using native-like discourse strategies.
High-Superior	This rating, which is not part of the ACTFL scale, is used in PST scoring for examinees who clearly exceed the requirements for a rating of Superior. A rating of High-Superior corresponds to a rating of 3+ to 5 on the scale used by the Interagency Language Roundtable of the U.S. Government. The PST is not designed to evaluate examinees above the ACTFL Superior level.

BEST COPY AVAILABLE

SOUTHEAST ASIAN
LANGUAGES PROFICIENCY EXAMINATIONS

James Dean Brown
H. Gary Cook
Charles Lockhart
Teresita Ramos

ABSTRACT

This paper reports on the design, administration, revision and validation of the Southeast Asian Summer Studies Institute (SEASSI) Proficiency Examinations. The goal was to develop parallel language proficiency examinations in each of five languages taught in the SEASSI: Indonesian, Khmer, Tagalog, Thai and Vietnamese. Four tests were developed for each of these languages: multiple-choice listening, interview, dictation and cloze test. To maximize the relationships among these examinations and the associated curricula, the interview and listening tests were each designed to assess all of the levels of language ability which are described in the *ACTFL Proficiency Guidelines* from "novice" to "advanced-plus."

This study (N = 218) explored the score distributions for each test on the proficiency batteries for each language, as well as differences between the distributions for the pilot (1989) and revised (1989) versions. The relative reliability estimates of the pilot and revised versions were also compared as were the various relationships among tests across languages.

The results are discussed in terms of the degree to which the scores on the strategies here are generalizable to test development projects for other Southeast Asian languages.

Each year since 1984, a Southeast Asian Summer Studies Institute (SEASSI) has been held on some university campus in the United States. As the name implies, the purpose of SEASSI is to provide instruction in the "lesser taught" languages from Southeast Asia. In 1988, SEASSI came to the university of Hawaii at Manoa for two consecutive summers. Since we found ourselves with several language testing specialists, a strong Indo-Pacific Language department, and two consecutive years to work, we were in a unique position to develop overall proficiency tests for a number of the languages taught in SEASSI -- tests that could then be passed on to future SEASSIs.

The central purpose of this paper is to describe the design, production, administration, piloting, revision and validation of these Southeast Asian Summer Studies Institute Proficiency Examinations (SEASSI). From the outset, the goal of this project was to develop overall language proficiency examinations in each of five languages taught in the SEASSI: Indonesia, Khmer, Tagalog, Thai and Vietnamese. The ultimate objectives of these tests was to assess the grammatical and communicative ability of students studying these languages in order to gauge their overall proficiency in the languages. It was decided early that the tests should be designed to measure all of the levels of language ability which are described in the ***ACTFL Proficiency Guidelines*** from "novice" to "advanced-plus" for speaking and listening (see Appendix A from ACTFL 1986, Liskin-Gasparro 1982, and/or ILR 1982). Though the ACTFL guidelines are somewhat controversial (eg. see Savignon 1985; Bachman and Savignon 1986), they provided a relatively simple paradigm within which we could develop and describe these tests in terms familiar to all of the teachers involved in the project, as well as to any language teachers who might be required to use the tests in the future.

The central research questions investigated in this were as follows

(1) How are the scores distributed for each test of the proficiency battery for each language, and how do the distributions differ between the pilot (1989) and revised (1989) versions?

(2) To what degree are the tests reliable? How does the reliability differ between the pilot and revised versions?

(3) To what degree are the tests intercorrelated? How do these correlation coefficients differ between the pilot and revised versions?

(4) To what degree are the tests parallel across languages?

(5) To what degree are the tests valid for purposes of testing overall proficiency in these languages?

(6) To what degree are the strategies described here generalizable to test development projects for other languages?

METHOD

A test development project like this has many facets. In order to facilitate the description and explanation of the project, this **METHOD** section will be organized into a description of the subject used for norming the tests, a section on the materials involved in the testing, an explanation of the procedures of the statistical procedures used to analyze, improve and reanalyze the tests.

Subject

A total of 228 students were involved in this project: 101 in the pilot stage of this project and 117 in the validation stage.

The 101 students involved in the pilot stage were all students in the SEASSI program during the summer of 1989 at the University of Hawaii at Manoa. They were enrolled in the first year (45.5%), second year (32.7%) and third year (21.8%) language courses in Indonesian (n = 26), Khmer (n = 21), Tagalog (n = 14) Thai (n = 17) and Vietnamese (n = 23). There were 48 females (47.5%) and 53 Males (52.5%). The vast majority of these students were native speakers of English (80.7%), though there were speakers of other languages who participated (19.3%).

The 117 students involved in the validation stage of this test development project were all students in the SEASSI program during summer 1989. They were enrolled in the first year (48.7%), second year (41.0%) and third year (10.3%) language courses in Indonesian (n = 54), Khmer (n = 18), Tagalog (n = 10) Thai (n = 23) and Vietnamese (n = 12). There were 57 females (48.7%) and 60 males (51.3%).

In general, all of the groups in this study were intact classes. To some degree, the participation of the students depended on the cooperation of their teachers. Since that cooperation was not universal, the samples in this project can only be viewed as typical of volunteer groups drawn from a summer intensive language study situation like that in SEASSI.

Materials

There were two test batteries employed in this project. The test of focus was the SEASSIPE. However, the ***Modern Language Aptitude Test*** (MLAT), developed by Carroll and Sapon (1959), was also administered. Each will be described in turn.

Description of the SEASSIPE. The SEASSIPE battery for each language presently consisted of four tests : multiple-choice listening, oral interview procedure, dictation and cloze test. In order to make the tests as comparable as possible across the five languages, they were all developed first in an English prototype version. The English version was then translated into the target language with an emphasis on truly translating the material into that language such that the result would be natural Indonesian, Khmer, Tagalog, Thai or Vietnamese. The multiple-choice ***listening*** test presented the students with aural statements or questions in the target language, and they were then asked what they would say (given four responses to choose from). The pilot versions of the test all contained 36 items, which were developed in 1988 on the basis of the ACTFL guidelines for listening (see **APPENDIX A**). The tests were then administered in the 1988 SEASSI. During 1989, the items were revised using distractor efficiency analysis, and six items were eliminated on the basis of overall item statistics. Thus the revised versions of the listening test all contained a total of 30 items.

The ***oral interview*** procedure was designed such that the interviewer would ask students questions at various levels of difficulty in the target language (based on the ACTFL speaking and listening guidelines in **APPENDIX A**). The students were required to respond in the target language. In the pilot version of the test, the responses of the students were rated on a 0-108 scale. On each of 36 questions, this scale had 0 to 3 points (one each for three categories: accuracy, fluency, and meaning). On the revised version of the interview, 12 questions were eliminated. Hence on the revised version, the students were rated on a 0-72 scale including one point each for accuracy, fluency and meaning based on a total of 24 interview questions.

The ***dictation*** consisted of an eighty word passage in the target language. The original English prototype was of approximately 7th grade reading level (using the ***Fry*** 1976 scale). The passage was read three times (once at normal rate of speech, then again with pauses at the end of logical phrases, and finally, again at normal rate). Each word that was morphologically correct was scored as a right answer. Because these dictations appeared to be working reasonably well, only very minor changes were made between the pilot and revised versions of this test.

The ***cloze*** test was based on an English prototype of 450 words at about the 7th grade reading level (again using the Fry 1976 scale). The cloze passage was created in the target language by translating the English passage and deleting every 13th word for a total of 30 blanks. The pilot and revised versions of this test each had the same number of items. However, blanks that proved ineffective statistically or linguistically in the pilot versions were changed to more promising positions in the revised tests (see Brown 1988b for more on cloze test improvement strategies).

As mentioned above, these four tests were developed for each of five languages taught in the SEASSI. To the degree that it was possible, they were made parallel across languages. The goal was that scores should be comparable across languages so that, for instance, a score of 50 on the interview procedure for Tagalog would be approximately the same as a score of 50 on the Thai test. To investigate the degree to which the tests were approximately equivalent across languages, the ***Modem Language Aptitude Test*** was also administered at the beginning of the instruction so that the results could be used to control for initial differences in language aptitude among the language groups.

All of the results of the SEASSI Proficiency Educations were considered experimental. Hence the results of the pilot project were used primarily to improve the tests and administration procedures in a revised version of each test. The scores were reported to the teachers to help in instructing and grading the students. However, the teachers were not required, in any way, to use the results, and the results were NOT used to judge the effectiveness of instruction. Teachers' input was solicited and used at all points in the test development process.

Description of the MLAT. The short version of the MLAT was also administered in this study. Only the last three of the five tests were administered as prescribed for the short version by the original authors. These three tests are entitled ***spelling clues, words in sentences and paired associates.***

The MLAT was included to control for differences in language learning aptitude across the five language groups and thereby help in investigating the equivalency of the tests across languages. The MLAT is a well-known language aptitude test. It was designed to predict performance in foreign language classroom. In this study, the results were kept confidential and did not affect the students' grades in any way. The scores and national percentile ranking were reported individually to the students with the caution that such scores represent only one type of information about their aptitude for learning foreign languages. It was made clear that the MLAT does not measure achievement in a specific language. The group scores, coded under anonymous student numbers, were only used to make general observations and to calculate some of the statistical analyses reported below.

Procedures

The overall plan for this project proceeded on schedule in four main stages and a number of smaller steps.

Stage one: Design. The tests were designed during June 1988 at the University of Hawaii at Manoa by J D Brown, Charles Lockhart and Teresita Ramos with the cooperation of teachers of the five languages involved (both in the Indo-Pacific Languages department and in SEASSI). J D Brown and C Lockhart were responsible for producing a prototypes into each of the five languages. J D Brown took primary responsibility for overall test design, administration and analysis.

Stage two: Production. The actual production of the tapes, booklets, answer sheets, scoring protocols and proctor instructions took place during the last week of July 1988 and the tests were actually administered in SEASSI classes on August 5, 1988. This stage was the responsibility of T. Ramos with the help of C. Lockhart.

Stage three: Validation. The on-going validation process involved the collection and organization of the August 5th data, as well as teacher ratings of the students' proficiency on the interview. Item analysis, descriptive statistics, correlational analysis and feedback from the teachers and students were all used to revise the four tests with the goal of improving them in terms of central tendency,dispersion, reliability and validity. The actual revisions and production of new versions of the tests took place during the spring and summer of 1989. This stage was primarily the responsibility of J D Brown with the help and cooperation of H Gary Cook, T Ramoa and the SEASSI teachers.

Stage four: Final Product. Revised versions of these tests were administered again in the 1989 SEASSI. This was primarily the job of H G Cook. A test manual was also produced (Brown, Cook, LocKhart and Ramos, unpublished ms). Based on the students' SEASSI performances and MLAT scores from both the 1988 and 1989 SEASSI, the manual provides directions for administering the tests, as well as discussion of the test development and norming procedures. The discussion focuses on the value of these new measures as indirect tests of ACTFL proficiency levels. The manual was developed following the standards set by AERA, APA and NCME in ***Standards for Educational and Psychological Testing*** (see APA 1985). The production of all tests, answer keys, audio tapes, answer sheets, manuals and reports was the primary responsibility of J D Brown.

Analyses

The analyses for this study were conducted using the *QuattroPro* spreadsheet program (Borland 1989), as well as the *ABSTAT* (Bell-Anderson 1989), and *SYSTAT* (Wilkinson, 1988) statistical program. These analyses fall into four categories: descriptive statistics, reliability statistics, correlational analyses, and analysis of covariance.

Because of the number of tests involved when we analyzed four tests each in two versions (1988 pilot version and 1989 revised version) for each of five languages (4 x 5 x 2 = 40), the ***descriptive statistics*** reported here are limited to the number of items, the number of subjects, the mean and the standard deviation. Similarity, ***reliability statistics*** have been limited to the Cronbach alpha coefficient (see Cronbach 1970) and the Kuder and Richardson (1973) formula 21 (K-R21). All ***correlation coefficients*** reported here are Pearson product-moment coefficients. Finally, ***analysis of covariance*** (ANCOVA) and multivariate analyses were used to determine the degree while controlling for differences in initial language aptitude (as measured by the MLAT). the alpha significance level for all statistical decisions was set at .05.

RESULTS

Summary descriptive statistics are presented in Table 1 for the pilot and revised versions of the four tests for each of the five languages. The languages are listed across the top of the table with the mean and standard deviation for each given directly below the language headings. The mean provides an indication of the overall central tendency, or typical behavior of a group, and the standard deviation gives an estimate of the average distance of students from the mean (see Brown 1988a for more on such statistics). The versions (ie. the pilot versions administered in summer of 1988 or the revised versions administered in summer of 1989) and tests (Listening, Oral Interview, Dictation and Cloze Test) are labeled down the left side of the table along with the number of items (k) in parentheses.

TABLE 1: MEASURE DESCRIPTIVE STATISTICS

VERSION	INDONESIAN		KHMER		TAGALOG		THAI		VIETNAMESE	
Test	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
PILOT 1988										
Listening (k = 32)	22.55	4.14	17.95	5.12	20.30	6.75	13.76	5.61	17.87	5.25
Oral Int. (k = 136)	77.65	2[illegible].50	77.64	23.56	63.92	20.62	63.27	9.72	74.08	18.1[illegible]
Dictation (k = 80)	[illegible].67	[illegible].12	35.[illegible]	[illegible]	33.87	1[illegible]	17.38	9.66	29.67	12.75
Cloze Test (k = 30)	17.12	*4.97	9.[illegible]0	4.06	12.23	5.75	16.50	3.07	15.64	6.57
REVISED 1989										
Listening (k = 30)	20.38	7.31	17.72	5.20	17.70	7.77	11.77	4.57	20.75	3.96
Oral Int. (k = 72)	64.59	6.66	61.89	9.55	56.50	10.32	50.72	13.92	57.33	6.31
Dictation (k = 80)	49.96	9.70	53.20	14.77	49.90	22.83	36.86	9.10	68.58	7.03
Cloze Test (k = 30)	20.09	4.12	17.90	4.15	17.90	6.33	13.50	4.34	14.50	5.30

Notice that, for each test, there is considerable variation across versions and languages not only in the magnitude of the means but also among the standard deviations. It seems probable that the disparities across versions (1988 and 1989) are largely due to the revision processes, but they may in part be caused by differences in the numbers of students at each level of study or by other differences among the samples used during the two summers.

Table 2 presents the reliabilities for each test based on the scores produced by the groups of students studying each of the languages. A reliability coefficient estimates the degree to which a test is consistent in what it measures. Such coefficient can range from 0.00 (wholly unreliable, or inconsistent) to 1.00 (completely reliable, or 100 percent consistent), and can take on all of the values in between, as well.

Notice that, once again, the languages are shown across the top of the table with two types of reliability, alpha and k-R21, labeled just under each language heading. You will also find that the versions (1988 or 1989) and tests are again labeled down the left side of the table.

TABLE 2: SEASSIPE TEST RELIABILITY FOR EACH LANGUAGE

VERSION	INDONESIAN		KHMER		TAGALOG		THAI		VIETNAMESE	
TEST	ALPHA	K-R21	ALPHA	K-R21	ALPHA	K-R21	ALPHA	K-R21	ALPHA	K-R21
[illegible] 1988										
Listening	[illegible]	.63	[illegible]	[illegible]	[illegible]	.86	[illegible]	.78	[illegible]	[illegible]
Oral Intv	[illegible]	.96	[illegible]	[illegible]	[illegible]	[illegible]	.75	.76	[illegible]	[illegible]
Dictation	**	.80	**	.89	**	.91	**	.50	**	[illegible]
Cloze Tst	*	.72	*	.64	*	.81	*	.22	*	.86
REVISED 1989										
Listening	[illegible]	.42	[illegible]	.76	.69	.51	.81	.68	.76	.61
Oral Intv	.91	.86	.54	.92	.95	.96	.97	.93	.78	.72
Dictation	**	.81	**	.92	**	.98	**	.78	**	.91
Cloze Tst	[illegible]	.63	.97	.60	.96	.85	.99	.63	.84	.76

* Not calculated.
** Not applicable.

As mentioned above, the reliability estimates reported in Table 2 are based on Cronbach alpha and on the K-R21. Cronbach alpha is an algebraic identity with the more familiar K-R20 for any test which is dichotomously scored (eg. the listening and cloze tests in this study). However, for any test which has a weighed scoring system (like the Interview tests in this study), another version of alpha must be applied -- in this case, one based on the odd-even variances (see Cronbach 1970)

TABLE 3: SEASSIPE TEST INTERCORRELATIONS FOR EACH LANGUAGE

VERSION	INDONESIAN			KHMER			TAGALOG			THAI			VIETNAMESE		
TEST(S)	L	O	D	L	O	D	L	O	D	L	O	D	L	O	D
PILOT 1988															
Oral Intv	54*			64			83*			.26			.65*		
Dictation	60*	78*		60*	.46*		72*	83*		-.12	.42		54*	.62*	
Cloze Tst	57*	79*	.82*	80	37*	.14*	.63*	.55*	.67*	-.24	.20	.94*	73*	.75*	.67*
REVISED 1989															
Oral Intv	51*			.70*			.[illegible]			.41			.59*		
Dictation	.38*	58*		.63*	.64*		.56	.83*		-.69	-.71		82*	.77*	
Cloze Tst	.30	.24	.58*	.64	.74*	.81*	.69*	.72*	[illegible]	56	.58	-.19	.73*	.77*	.72*

* p < .05
L = LISTENING; O = ORAL INTV; D = CLOZE TST

Intercorrelations among the SEASSIPE tests on both versions were calculated using the Pearson product-moment correlation coefficient for each language separately (see Table 3). A correlation coefficient gives an estimate of the degree to which two sets of numbers are related. A coefficient of 0.00 indicates that the numbers are totally unrelated. A coefficient of +1.00 indicates that they are completely related (mostly in terms of being ordered in the same way). A coefficient of -1.00 indicates that they are strongly related, but in opposite directions, ie. as one set of numbers becomes larger, the other set grows smaller. Naturally, coefficients can vary throughout this range from -1.00 to 0.00 to 1.00.

Notice that the languages are labeled across the top with Listening (L), Oral Interview (O) and Dictation (D) also indicated for each language. The versions (1988 or 1989) and tests (Oral Interview, Dictation and Cloze Test) are also indicated down the left side. To read the table, remember that each correlation coefficient is found at the intersection of the two variables that were being examined. This means, for instance, that the .54 in the upper-left corner indicates the degree of relationship between the scores on the Oral Interview and Listening tests in Indonesian in 1988 pilot version.

Following some of the correlation coefficients in Table 3, there is an asterisk, which refers down below the table to p < .05. This simply means that these correlation coefficients are statistically significant at the .05 level. In other words, there is only a five percent probability that the correlation coefficients with asterisks occurred by chance alone. Put yet another way, there is a 95 percent probability that the coefficients with asterisks occurred for other than chance reasons. Those coefficients without asterisks can be interpreted as being zero.

Recall that, in Table 1, there was considerable variation in the magnitude of the means and standard deviations across languages and versions. Table 4 shows the results of an analysis of covariance procedure which used language (Indonesian, Khmer, Tagalog, Thai and Vietnamese) as a categorical variable and MLAT language aptitude scores as a covariate to determine whether there were significant differences across languages for the mean test scores (Listening, Interview, Dictation and Cloze treated as repeated measures).

TABLE 4: ANALYSIS OF COVARIANCE ACROSS REPEATED MEASURES (TESTS)

SOURCE	SS	df	MS	F
BETWEEN SUBJECTS				
LANGUAGE	3197.197	4	799.299	7.248*
MLAT (COVARIATE)	256.014	1	256.014	2.322
SUBJECTS WITHIN GROUPS	6285.642	57	110.274	
WITHIN SUBJECTS				
LANGUAGE	7156.650	12	596.387	18.196*
MLAT (COVARIATE)	80.513	3	26.838	0.819
SUBJECTS WITHIN GROUPS	5604.643	171	32.776	

*p < .05

In Table 4, it is important to realize that the asterisks next to the F ratios indicate that there is some significant difference among the means for different languages across the four tests. This means in effect that at least one of the differences in means shown in table 1 is due to other than chance factors (with 95 percent certainly). Of course, many more of the differences may also be significant, but there is no way of knowing which they are from this overall analysis. It should suffice to recognize that a significant difference exists somewhere across languages. The lack of asterisks after the F ratios for the MLAT indicate that there was no significant difference detected language aptitude (as measured by MLAT) among the groups of students taking the five languages.

Since analysis of covariance is a fairly controversial procedure, two additional steps were taken:

(1) First, the assumption of homogeneity of slopes was carefully checked by calculating and examining the interaction terms before performing the actual analysis of covariance. The interactions were not found to be significant.

(2) Second, multivariate analyses (including, Wilks' lambda, Pillai trace, and Hotelling-Lawley trace) were also calculated. Since they led to exactly the same conclusions as the univariate statistics shown in Table 4, they are not reported here.

Thus the assumptions were found to be met for the univariate analysis of covariance procedures in a repeated measures design, and the results were further confirmed using multivariate procedures. It is therefore with a fair amount of confidence that these results are reported here.

TABLE 5: DIFFERENTIAL PERFORMANCE BY LEVELS ON EACH TEST

TEST	LEVEL	MEAN	STD	N
Listening	1st year	15.7347	5.2392	49
	2nd year	19.6383	4.4007	47
	3rd year	20.9167	4.5218	12
Oral Intv	1st year	50.6538	14.9022	26
	2nd year	47.1915	15.9519	47
	3rd year	57.5000	12.2734	12
Dictation	1st year	16.4063	5.3573	32
	2nd year	18.2500	4.2602	48
	3rd year	23.9167	3.4499	12
Cloze Tst	1st year	57.3393	12.1015	56
	2nd year	61.5000	8.8894	48
	3rd year	65.5833	6.8948	12

One other important result was found in this study: the tests do appear to reflect the differences in ability found between levels of language study. This is an important issue for overall proficiency tests like the SEASSIPE because they should be sensitive to the types of overall differences in language ability that would develop over time, or among individuals studying at different levels. While this differential level effect was found for each of the languages, it is summarized across languages in Table 5 (in the interests of economy of space). Notice that, with one exception, the means get higher on all of the tests as the

level of the students goes up from first to second to third year. The one anomaly is between the first and second years on the oral interview.

DISCUSSION

The purpose of this section will be to interpret the results reported above with the goal of providing direct answers to the original research questions posed at the beginning of this study. Consequently, the research questions will be restated and used as headings to help organize the discussion.

(1) *How are the scores distributed for each test of the proficiency battery for each language, and how do the distributions differ between the pilot (1989) and revised (1989) versions?*

The results in Table 1 indicate that most of the current tests are reasonably well-centered and have scores that are fairly widely dispersed about the central tendency. Several notable exceptions seem to be the 1989 Oral Interviews for Indonesian and Khmer, both of which appear to be negatively skewed (providing classic examples of what is commonly called the ceiling effect -- see Brown 1988a for further explanation). It is difficult, if not impossible, to disentangle whether the differences found between the two versions of the test (1988 and 1989) are due to the revision processes in which many of the tests were shortened and improved, or to differences in the samples used during the two SEASSIs.

(2) *To what degree are the tests reliable? How does the reliability differ between the pilot and revised versions?*

Table 2 shows an array of reliability coefficients for the 1988 pilot version and 1989 revised tests that are all moderate to very high in magnitude. The lowest of these is for the 1989 Indonesian Listening test. It is low enough that the results for this test should only be used with extreme caution until it can be administered again to determine whether the low reliability is a result of bad test design or some aspect of the sample of students who took the test.

These reliability statistics indicate that most of the tests produce reasonably consistent results even when they are administered to the relatively homogeneous population of SEASSI students. The revision process appears to have generally, though not universally, improved test reliability either in terms of producing higher reliability indices or approximately equal estimates, but for shorter more efficient, versions. The listening tests for Indonesian and Tagalog are worrisome because the reliabilities are lower in the revised than in the pilot testing and because they are found among the 1989 results. However, it is important to remember that these are fairly short tests and that they are being administered to relatively restricted ranges of ability in the various languages involved. These are both important factors because, all things being equal, a short test will be less reliable than a long test, and a restricted range of talent will produce lower reliability estimates than a wide one (for further explanation and examples, see Ebel 1979; Brown 1984, 1988a).

Note also that the K-R21 statistic is generally lower than the alpha estimate. This is typical. K-R21 is a relatively easy to calculate reliability estimate, but it usually underestimates the actual reliability of the test (see, for instance, the 1989 Revis. i Khmer and Thai cloze tests reliabilities in Table 2).

(3) *To what degree are the tests intercorrelated? How do these correlation coefficients differ between the pilot and revised versions?*

In most cases, the correlation coefficients reported in Table 3 indicate a surprisingly high degree of relationship among the tests. The one systematic and glaring exception is the set of coefficients found for Thai. It is important to note that these correlation coefficients for Thai based on very small samples (due mostly to the fact that students at the lowest level were not taught to write in Thai), and that these correlation coefficients were not statistically significant at the $p < .05$ level. They must therefore be interpreted as correlation coefficients that probably occurred by chance alone, or simply as correlations of zero.

(4) *To what degree are the tests parallel across languages?*

The interpretation of these results is fairly straightforward. Apparently, there was no statistically significant difference in MLAT language aptitude scores among the groups studying the five languages. However, there was clearly a significant difference among the mean test scores across the five languages despite the efforts to control for initial differences in language aptitude (the MLAT covariate). A glance back at Table 1 will indicate the magnitude of such differences.

One possible cause for these differences is that the tests have changed during the process of development. Recall that all of these tests started out as the same English language prototype . It is apparent that, during the processes of translating and revising, the tests diverged in overall difficulty across languages. this is reflected in the mean differences found here. Another potential cause of the statistically significant differences reported in Tables 1, 4, and 5 is that there may have been considerable variations in the samples used during the two summers.

(5) *To what degree are the tests valid for purposes of testing overall proficiency in these languages?*

The intercorrelations among the tests for each language (see Table 3) indicate that moderate to strong systematic relationships exist among many of the tests in four of the five languages being tested in this project (the exception is Thai). However, this type of correlational analysis is far from sufficient for analysing the validity of these tests. If there were other well established tests of the skills being tested in these languages, it would be possible to administer those criterion tests along with the SEASSIPE tests and study the correlation coefficients between our relatively new tests and the well-established measures. Such information could then be used to build arguments for the criterion-related validity of some or all of these measures. Unfortunately, no such well-established criterion measures were available at the time of this project.

However, there are results in this study that do lend support to the construct validity of these tests. The fact that the tests generally reflect differences between levels of study (as shown in Table 3) provides evidence for the construct validity (the differential groups type) of these tests.

Nevertheless, much more evidence should be gathered on the validity of the various measures in this study. An intervention study of their construct validity could be set up by administering the tests before and after instruction to determine the degree to which they are sensitive to the language proficiency construct which is presumably being taught in the course. If, in future data, correlational analyses indicate patterns similar to those found here, factor analyses factor analysis might also be used profitably to explore the variance structures of those relationships.

The point is that there are indications in this study of the validity of the tests involved. However, in the study of validity, it is important to build arguments from a number of perspectives on an ongoing basis. Hence, in a sense, the study of validity is never fully complete as long as more evidence can be gathered and stronger arguments can be constructed.

(6) *To what degree are the strategies described here generalizable to test development projects for other languages?*

From the outset, this project was designed to provide four different types of proficiency tests -- tests that would be comparable across five languages. The intention was to develop tests that would produce scores that were comparable across languages such that a score of 34 would be roughly comparable in Indonesian, Khmer, Tagalog, Thai and Vietnamese. Perhaps this entire aspect of the project was quixotic from the very beginning. Recall that the process began with the creation of English language prototypes for the listening test, oral interview, dictation and cloze procedure. These prototypes were then translated into the five languages with strict instructions to really translate them, ie. to make them comfortably and wholly Indonesian, Khmer, Tagalog, Thai and Vietnamese. While the very act of translating the passages in five different directions probably affected their comparability across languages, they probably remained at least roughly the same at this stage of development. Then, during the summer of 1988, the tests were administered, analyzed and revised separately using different samples of students with the result that the tests further diverged in content and function.

We now know that the use of English language prototypes for the development of these tests may have created problems that we did not foresee. One danger is that such a strategy avoids the use of language that is authentic in the target language. For instance, a passage that is translated from English for use in Khmer cloze test may be topic that would never be discussed in the target culture, may be organized in a manner totally alien to Khmer, or may simply seem stilted to native speakers of Khmer because of its rhetorical structure. These problems could occur no matter how well-translated the passage might be.

Ultimately, the tests did not turn out to be similar enough across languages to justify using this translation strategy. Thus we do not recommend its use in further test development projects. It would probably have been far more profitable to use authentic materials from the countries involved to develop tests directly related to the target languages and cultures.

CONCLUSION

In summary, the tests in each of the five SEASSI Proficiency Examinations appear to be reasonably well-centered and seem to adequately disperse the students' performance. They are also reasonably reliable. Naturally, future research should focus on ways to make the tests increasingly reliable and further build a case for their validity. Thus the final versions of the tests can be passed on to future SEASSIs at other sites with some confidence that any decisions based on them will be reasonably professional and sound. It is also with some confidence that the tests will be used here at the University of Hawaii at Manoa to test the overall proficiency of students studying Indonesian, Khmer, Tagalog, Thai and Vietnamese. However, the process of test development and revision should never be viewed as finished. Any test can be further improved and made to better serve the population of students and teachers who are the ultimate users of such materials.

One final point must be stressed: we could never have successfully carried out this project without the cooperation of the many language teachers who volunteered their time while carrying out other duties in the Indo-Pacific Languages department, or the SEASSIs held at University of Hawaii at Manoa. We owe each of these language teachers a personal debt of gratitude. Unfortunately, we can only thank them as a group for their professionalism and hard work.

REFERENCES

ACTFL (1986). **ACTFL proficiency guidelines.** Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Anderson-Bell. (1989). **ABSTAT.** Parker, CO: Anderson-Bell.

APA. (1985). **Standards for educational and psychological testing.** Washington, DC: American Psychological Association.

Bachman, L & S Savignon. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. **Modern Language Journal,** 70, 380-397.

Borland. (1989). **QuattroPro.** Scotts Valley, CA: Borland International.

Brown, J D. (1983). A closer look at cloze: validity and reliability. In J W Oller, Jr (Ed). **Issues in language testing.** Cambridge, MA: Newbury House.

Brown, J D. (1984). A cloze is a cloze is a cloze? In J Handscombe, R A Orem, and B P Taylor (Eds). **On TESOL '83: the question of control.** Washington, DC: TESOL.

Brown, J D. (1988a). **Understanding research in second language learning: A teacher's guide to statistics and research design.** London: Cambridge University.

Brown, J D. (1988b). Tailored cloze: improved with classical item analysis techniques. **Language Testing,** 5, 19-31.

Brown, J D., H G Cook, C Lockhart and T Ramos (Unpublished ms). The SEASSI Proficiency Examination Technical Manual. Honolulu, HI: University of Hawaii at Manoa.

Carroll, J B and S M Sapon (1959). **Modem language aptitude test.** New York: The Psychological Corporation.

Cronbach, L J. (1970). **Essentials of psychological testing** (3rd ed). New York: Harper and Row.

Ebel, R L. (1979). **Essentials of educational measurement** (3rd ed). Englewood Cliffs, NJ: Prentice-Hall.

Fry, E B. (1976). **Fry readability scale (extended).** Providence, RI: Jamestown Publishers.

ILR (1982). Interagency Language Roundtable Language Skill Level Descriptions: Speaking. appendix B in Liskin-Gasparro (1982).

Kuder, G F and M W Richardson. (1937). The theory of estimation of test reliability. **Psychometrika,** 2, 151-160.

Liskin-Gasparro (1982). **Testing and teaching for oral proficiency.** Boston: Heinle and Heinle.

Savignon, S J. (1985). Evaluation of communicative competence: the ACTFL provisional proficiency guidelines. **Modem Language Journal,** 69, 129-142.

Wilkinson, L. (1988). **SYSTAT: The system for statistics.** Evanston, IL: SYSTAT.

APPENDIX A
ACTFL PROFICIENCY GUIDELINES FOR SPEAKING AND LISTENING
(ACTFL 1986)

Generic Descriptions-Speaking

Novice

The Novice level is characterized by the ability to communicate minimally with learned material

Novice Low

Oral production consists of isolated words and perhaps a few high-frequency phrases Essentially no functional communicative ability

Novice Mid

Oral production continues to consist of isolated words and learned phrases within very predictable areas of need, although quantity is increased. Vocabulary is sufficient only for handling simple, elementary needs and expressing basic courtesies Utterances rarely consist of more than two or three words and show frequent long pauses and repetition of interlocutor's words. Speaker may have some difficulty producing even the simplest utterances. Some Novice-Mid speakers will be understood only with great difficulty

Novice High

Able to satisfy partially the requirements of basic communicative exchanges by relying heavily on learned utterances but occasionally expanding these through simple recombinations of their elements. Can ask questions or make statements involving learned material Shows signs of spontaneity although this falls short of real autonomy of expression Speech continues to consist of learned utterances rather than of personalized, situationally adapted ones Vocabulary centers on areas such as basic objects, places, and most common kinship terms. Pronunciation may still be strongly influenced by first language. Errors are frequent and, in spite of repetition, some Novice High speakers will have difficulty being understood even by sympathetic interlocutors

Intermediate

The Intermediate level is charaeterized by the speaker's ability to:
—create with the language by combining and recombining learned elements, though primarily in a reactive mode
—initiate, minimally sustain, and close in a simple way basic communicative tasks, and
—ask and answer questions.

Intermediate-Low

Able to handle successfully a limited number of interactive, task-oriented and social situations Can ask and answer questions, initiate and respond to simple statements and maintain face to-face conversation, although in a highly restricted manner and with much linguistic inaccuracy. Within these limitations, can perform such tasks as introducing self, ordering a meal, asking directions, and making purchases. Vocabulary is adequate to express only the most elementary needs. Strong interference from native language may occur Misunderstandings frequently arise, but with repetition, the Intermediate-Low speaker can generally be understood by sympathetic interlocutors.

Intermediate-Mid

Able to handle successfully a variety of uncomplicated, basic and communicative tasks and social situations Can talk simply about self and family members Can ask and answer questions and participate in simple conversations on topics beyond the most immediate needs; e g., personal history and leisure time activities Utterance length increases slightly, but speech may continue to be characterized by frequent long pauses, since the smooth incorporation of even basic conversational strategies is often hindered as the speaker struggles to create appropriate language forms Pronunciation may continue to be strongly influenced by first language and fluency may still be strained. Although misunderstandings still arise, the Intermediate-Mid speaker can generally be understood by sympathetic interloculors

Intermediate High

Able to handle successfully most uncomplicated communicative tasks and social situations Can initiate, sustain, and close a general conversation with a number of strategies appropriate to a range of circumstances and topics, but errors are evident. Limited vocabulary still necessitates hesitation and may bring about slightly unexpected circumlocution. There is emerging evidence of connected discourse, particularly for simple narration and/or description. The Intermediate High speaker can generally by understood even by interlocutors not accustomed to dealing with speakers at this level, but repetition may still be required.

Advanced The Advanced level is characterized by the speaker's ability to:

—converse in a clearly participatory fashion:

—initiate, sustain, and bring to closure a wide variety of communicative tasks, including those that require an increased ability to convey meaning with diverse language strategies due to a complication or an unforeseen turn of events,

—satisfy the requirements of school and work situations; and

—narrate and describe with paragraph-length connected discourse.

Advanced Able to satisfy the requirements of everyday situations and routine school and work requirements. Can handle with confidence but not with facility complicated tasks and social situations, such as elaborating, complaining, and apologizing. Can narrate and describe with some details, linking sentences together smoothly. Can communicate facts and talk casually about topics of current public and personal interest, using general vocabulary. Shortcomings can often be smoothed over by communicative strategies, such as pause fillers, stalling devices, and different rates of speech. Circumlocution which arises from vocabulary or syntactic limitations very often is quite successful, though some groping for words may still be evident. The Advanced level speaker can be understood without difficulty by native interlocutors

Advanced Plus Able to satisfy the requirements of a broad variety of everyday, school, and work situations. Can discuss concrete topics relating to particular interests and special fields of competence. There is emerging evidence of ability to support opinions, explain in detail, and hypothesize. The Advanced Plus speaker often shows a well developed ability to compensate for an imperfect grasp of some forms with confident use of communicative strategies, such as paraphrasing and circumlocution. Differentiated vocabulary and intonation are effectively used to communicate fine shades of meaning. The Advanced-Plus speaker often shows remarkable fluency and ease of speech but under the demands of Superior-level, complex tasks, language may break down or prove inadequate

Superior The Superior level is characterized by the speaker's ability to

—participate effectively in most formal and informal conversations on practical, social, professional, and abstract topics, and

—support opinions and hypothesize using native like discourse strategies

Superior Able to speak the language with sufficient accuracy to participate effectively in most formal and informal conversations on practical, social, professional, and abstract topics. Can discuss special fields of competence and interest with ease. Can support opinions and hypothesize, but may not be able to tailor language to audience or discuss in depth highly abstract or unfamiliar topics. Usually the Superior level speaker is only partially familiar with regional or other dialectical variants. The Superior level speaker commands a wide variety of interactive strategies and shows good awareness of discourse strategies. The latter involves the ability to distinguish main ideas from supporting information through syntactic, lexical and suprasegmental features (pitch, stress, intonation). Sporadic errors may occur, particularly in low-frequency structures and some complex high-frequency structures more common to formal writing, but no patterns of error are evident. Errors do not disturb the native speaker or interfere with communication

Generic Descriptions-Listening

These guidelines assume that all listening tasks take place in an authentic environment at a normal rate of speech using standard or near standard norms

Novice Low Understanding is limited to occasional isolated words, such as cognates, borrowed words, and high frequency social conventions. Essentially no ability to comprehend even short utterances

Novice Mid Able to understand some short, learned utterances, particularly where context strongly supports understanding and speech is clearly audible. Comprehends some words and phrases from simple questions, statements, high frequency commands and courtesy formulae about topics that refer to basic personal information or the immediate physical setting. The listener requires long pauses for assimilation and periodically requests repetition and/or a slower rate of speech

Novice-High

Able to understand short, learned utterances and some sentence-length utterances, particularly where context strongly supports understanding and speech is clearly audible. Comprehends words and phrases from simple questions, statements, high-frequency commands and courtesy formulae. May require repetition, rephrasing and/or a slowed rate of speech for comprehension.

Intermediate-Low

Able to understand sentence-length utterances which consist of recombinations of learned elements in a limited number of content areas, particularly if strongly supported by the situational context. Content refers to basic personal background and needs, social conventions and routine tasks, such as getting meals and receiving simple instructions and directions. Listening tasks pertain primarily to spontaneous face-to-face conversations. Understanding is often uneven; repetition and rewording may be necessary. Misunderstandings in both main ideas and details arise frequently.

Intermediate-Mid

Able to understand sentence-length utterances which consist of recombinations of learned utterances on a variety of topics. Content continues to refer primarily to basic personal background and needs, social conventions and somewhat more complex tasks, such as lodging, transportation, and shopping. Additional content areas include some personal interests and activities, and a greater diversity of instructions and directions. Listening tasks not only pertain to spontaneous face-to-face conversations but also to short routine telephone conversations and some deliberate speech, such as simple announcements and reports over the media. Understanding continues to be uneven.

Intermediate-High

Able to sustain understanding over longer stretches of connected discourse on a number of topics pertaining to different times and places; however, understanding is inconsistent due to failure to grasp main ideas and/or details. Thus, while topics do not differ significantly from those of an Advanced level listener, comprehension is less in quantity and poorer in quality.

Advanced

Able to understand main ideas and most details of connected discourse on a variety of topics beyond the immediacy of the situation. Comprehension may be uneven due to a variety of linguistic and extralinguistic factors, among which topic familiarity is very prominent. These texts frequently involve description and narration in different time frames or aspects, such as present, nonpast, habitual, or imperfective. Texts may include interviews, short lectures on familiar topics, and news items and reports primarily dealing with factual information. Listener is aware of cohesive devices but may not be able to use them to follow the sequence of thought in an oral text.

Advanced Plus

Able to understand the main ideas of most speech in a standard dialect; however, the listener may not be able to sustain comprehension in extended discourse which is propositionally and linguistically complex. Listener shows an emerging awareness of culturally implied meanings beyond the surface meanings of the text but may fail to grasp sociocultural nuances of the message.

Superior

Able to understand the main ideas of all speech in a standard dialect, including technical discussion in a field of specialization. Can follow the essentials of extended discourse which is propositionally and linguistically complex, as in academic/professional settings, in lectures, speeches, and reports. Listener shows some appreciation of aesthetic norms of target language, of idioms, colloquialisms, and register shifting. Able to make inferences within the cultural framework of the target language. Understanding is aided by an awareness of the underlying organizational structure of the oral text and includes sensitivity for its social and cultural references and its affective overtones. Rarely misunderstands but may not understand excessively rapid, highly colloquial speech or speech that has strong cultural references.

Distinguished

Able to understand all forms and styles of speech pertinent to personal, social and professional needs tailored to different audiences. Shows strong sensitivity to social and cultural references and aesthetic norms by processing language from within the cultural framework. Texts include theater plays, screen productions, editorials, symposia, academic debates, public policy statements, literary readings, and most jokes and puns. May have difficulty with some dialects and slang.

CONTINUOUS ASSESSMENT IN THE ORAL COMMUNICATION CLASS: TEACHER CONSTRUCTED TEST

Shanta Nair - Venugopal

INTRODUCTION

The Teacher As Tester

There is evidence that not only are teachers good judges of behaviour, they are also reliable judges of test performances. (Callaway, D R 1980). However it would be quite naive and perhaps even imprudent to suggest then, that all teachers will also by extension make naturally good testers given Spolsky's (1975) rhetoric on whether testing is art or science. Nevertheless, it can be assumed that a teacher who has been actively involved in course design or better still in the privileged position of 'negotiating' the curriculum, with her students would at least have a blueprint of sorts as a starting point for the construction of tests for that course. This could be further enhanced if the process is subjected to friendly criticism at the very least by other members of staff in relation to the objectives of the course or curriculum as a whole. The teacher is then in the informed and educated position of being able to translate the objectives of the course into tests construction by linking the specific objectives of the course with the task specifications identified. The test would then be underpinned by at least a view of language learning even if not a full fledged theory, in a clear case of doing the best that can be done. The analogy is best supplied by Skehan (1988) who summarized the current state of the art on (communicative) testing.

"...Since ... definitive theories do not exist, testers have to do the best they can with such theories as are available."

The contention therefore is that the teacher who has had some responsibility for course design and implementation is in many ways pre-eminently qualified to construct tests for the course particularly if it is backed by experience and shared knowledge in the field. Since the target group is known at first hand, needs can be fairly accurately specified on the basis of introspection and experience. The backwash effect of teacher-made tests on teaching can only be beneficial. As the teacher in this case is also responsible for course content (and like all other teachers across the board has the best interests of her students at heart), she will certainly teach what is to be tested, test what is taught and 'bias for best' in the use of test procedures and situations. The only possible danger lurking in this happy land is the possibility of a teacher who willy-nilly teaches the test as well and thereby nullifies its value as a measuring instrument.

BACKGROUND

The Target Group

At the English Department of the National University of Malaysia (UKM), students in the second year of the B A in English Studies program are required to take both levels 1 and 2 of an oral communication course that straddles two semesters or one academic session. These students are viewed as potential candidates for the B A in English Studies degree and there is a tremendous responsibility (equally shared by the writing and reading courses) to improve their language ability to make them "respectable" (Nair-Venugopal, S. 1988) candidates for the program. This may be seen as the perceived and immediate need. The projected or future need is seen as a high level of language ability that also makes for good language modelling as there is evidence that many of these students upon graduation enroll for a diploma in Education and become English language teachers. The mature students in the course are invariably teachers too. The responsibility is even more awesome given the language situation in the country which while overtly ESL also manifests many hybrids of the ESL/EFL situation, notwithstanding government efforts at promoting English as an important second language. These students (except those who are exempted on the basis of a placement test and have earned credits equivalent to the course) are also subject to a one year fairly intensive preparatory proficiency program (twelve hours per week). The emphasis in this course is on an integrated teaching of the four language skills. These students have also had a minimum of eleven years of instruction in English as a subject in school. There is also invariably the case of the mature student who has probably had 'more' English instruction, having been subject chronologically to a different system of education in the country's history.

Course Objectives

The oral communication course comprises two levels- each level taught over two semesters consecutively. The general aim of level 1 is to provide a language learning environment for the acquisition of advanced oral skills and that of level II to augment and improve upon the skills acquired in level I, thus providing a learning continuum for the acquisition of advanced oral skills. At this juncture it must be pointed out that in the integrated program of the first year there is an oral fluency component. In other words the students in the second year have already been thrown into the 'deep end' as it were and the assumption is that upon entry to Level I they have more than banal or survival skills in oral communication. The reality is that students in spite of the first year of fairly intensive instruction and exposure enter the second year with varying levels of abilities. The task at hand for the second year oral skills programme is quite clear; raise levels of individual oral ability, bridge varying levels of individual abilities and yet help students to develop at their own pace. Hence the need to see the language class as a language acquisition environment bearing in mind that contact and exposure with the language outside the class is not optimal. The main objective in Level I is to achieve a high level of oral fluency in the language with an accompanying level of confidence and intelligibility, the latter being viewed with some urgency since native vernaculars are increasingly used for social communication outside the classroom and Bahasa Malaysia remains the language of instruction for courses in all other disciplines. The main objective of Level II is to achieve a high level of oral language ability. Both these objectives are further broken down into specific objectives for both levels. The tests are pegged against these objectives.

The specific objectives of Level I of the course are as follows:

1 attain high levels of intelligibility in speech

2 comprehend standard varieties of the spoken language without difficulty

3 interact and converse freely among themselves and other speakers of the language

4 convey information,narrate and describe; express and justify opinions.

These objectives are realized through an eclectic methodology using a variety of instructional devices, classroom procedures and multimedia materials.

The second objective is realized largely through practice in the language laboratory and it is not tested ie. elicited for as a skill domain in the tests that have been developed for the course. While it is generally accepted that listening comprehension as a skill is not easy to teach, it is even more elusive to test. According to Brown, G. and Yule, G. (1983)

> "...a listener's task performance may be unreliable for a number of reasons... we have only a very limited understanding of how we could determine what it is that listening comprehension entails. Given these two observations, it woul·' seem that the assessment of listening comprehension is an extremely complex undertaking".

Having said that, why then has listening comprehension been included as a desirable objective on the course? As the view of language underlying the course is that of communication, no course that purports to teach oral communication (which view of language surely sees listening as a reciprocal skill) can justifiably not pay attention to teaching it at least. Objective 3 is specifically tested as speech interaction in the form of group discussions and 4 as extended "impromptu" speech in 3 modes. 1 is rated as a variable of performance for both these test types. 4 is also subsumed as 'enabling' skills in the group discussion test.

Objectives for level 2 are as follows:

1 not only comprehend all standard varieties of the language but also make themselves understood to other speakers of the language without difficulty.

2 participate in discussions on topics of a wide range of general interest without hesitation or effort

3 speak before audiences confidently (as in public speaking/platform activities)

4 convey information, persuade others and express themselves effectively as users of the language (as in debates and forums)

These objectives are achieved through the use of a selection of instructional devices, classroom procedures and modes such as simulations, small group discussions, debates and public speaking.

Objective 2 is tested using the group discussion test. 3 and 4 to borrow Tarone's notion (1982/83) of a "continuum of interlanguage styles" are to be seen as examples of "careful styles" and are tested as formal modes of speaking and debates. Objective 4 is also elicited as performance variables in the group discussion test. The second part of 1 ie. intelligibility/comprehensibility operates as an important variable in assessing the performance of all these tests. The final tests for both leve' sample global communicative ability in the rehearsed speech genre which is an oral newsmagazine presentation on tape for the first level and a videotaped presentation for the second level of either one of two platform activities or a chat show. Both are take-home, end-of-semester projects.

THE TESTS

Some Considerations

"In constructing tests, it is essential to have a defined curriculum or a set body of knowledge from which testers determine what to test (Shohamy, E 1988)".

To echo Charles Alderson (1983) the most important question to be asked of any test is, "What is it measuring?" which "can be determined by a variety of means including face inspection". Needless to say there are two other questions that merit equal consideration. One is, how is it measured and perhaps more crucially why? With reference to these tests, the question "for whom" ie. the target group has already been answered. As for purpose, each test type is seen as having a specified purpose that corresponds to an ability in an oral skill domain that has been delineated in the course objectives. Task specifications are prescribed by the oral skills domains. Therefore each test would sample different behaviour or skills in the form of different speech modes and the task specifications will vary from test type to test type. However all tests will test for both linguistic and communicative ability.

> "It is difficult to totally separate the two criteria, as the linguistic quality of an utterance can influence comprehensibility the basic communicative criterion. Further, while a major goal of most college or secondary language programs is communicative ability in the target language, there is justifiable concern with linguistic correctness because ...we are not just attempting to teach survival communications..., we are also trying to teach literacy in another language". Bartz W H (1979)

It is quite clear that as the view of the language underlying the teaching is communicative and the view of language learning, that of acquisition, achievement tests administered both mid-way and at the end of each semester will not allow the teacher to obtain feedback on acquired ability which could be used for diagnostic purposes as well (particularly at entry from the first level to the second), nor allow for a 'profiling' of performance. Hence the need for and the development of a continuous 'battery' of tests, spaced out in relation to their ordering on the course and as spelt out by the course objectives. These have been conceptualized as oral skills domains and rated accordingly.

> "...Advances in the state of the art of achievement testing are directly related to advances in the concept of skills domains on which student achievement is assessed". Shoemaker (cited by Swain M. 1980)

The tests are administered at various points in the semesters that roughly coincide with points on the course where the skills to be tested have already been taught or practised. The course provides ample opportunity in the practice of these sk'lls. Such an ordering on the learning continuum had implications for the content validity of the tests where,

> "Content validity refers to the ability of a test to measure what has been taught and subsequently learned by the students. It is obvious that teachers must see that the test is designed so that it contains items that correlate with the content of instruction. Thus it follows that unless students are given practice in oral communication in the foreign language classroom, evaluation of communication may not be valid...." Bartz (W H 1979).

By spacing out the tests in relation to the content, not only is the teacher-tester able to 'fit' the test to the content, she is also able after each test to obtain valuable feedback for the teaching of the subsequent domains that have been arranged in a cyclical fashion. Hence learning and performance is also on a cumulative basis because each skill taught and learnt or acquired presupposes and builds on the acquisition and the development of the preceding skills. It is on these bases that the tests have been developed and administered over a period of time. They are direct tests of performance that are communicative in nature and administered on a cumulative basis as part of on-going course assessment for both levels. The tests formats, and methods of elicitation owe much to some knowledge in the field (particularly the state of the art), test feedback, student introspection and teacher retrospection and experience with its full range of hunches and intuition.

Test Types

Level I

Level I as mentioned earlier consists of three test types.

1 Extended/'impromptu' speech

2 Group discussion

3 End-of-semester project

There are three speaking tasks of this type. Student speak for about 2 minutes on the first, 2-3 on the second and 3-5 on the third. The tasks test for three modes of speech as follows:

(i) Talking about oneself, others and experiences

(ii) Narrating and describing incidents and events

(iii) Expressing and justifying opinions.

1 (i) and (ii) are tested at the beginning of the first level mainly for diagnostic purposes as the students are of heterogeneous levels of proficiency. The speeches are staggered for both (i) and (iii) to ensure that each student has a minimum of a minute or so to prepare mentally for the topic. For (ii) they are all given an equal amount of time to prepare mentally and to make notes. When the testing begins they listen to each other speak, as the audience, thus providing the motivation and a 'valid' reason as it were for the task. (iii) is tested before the second half of the semester, to obtain information on learned behaviour as the students have had sufficient practice in expressing and justifying opinions through reaching consensus in group work. The topics for (i) and (ii) are well within the students' realm of experience and interest such as

The happiest day in my life.
The person who has influenced me the most.

However the topics for (iii) are of a slightly controversial nature such as

Should smoking be banned in all public places?
Do women make better teachers?

Both (ii) and (iii) are rated for global ability to communicate in the mode which is the overall ability of the student to persuade or justify reasons taken for a stand in the case of the latter and to describe, report and narrate in the case of the former.

2 The group discussion test is administered in the second half of the semester as by this time there has been plenty of practice in the interaction mode as the modus operandi of Level I is small group work. It tests specifically for oral interaction skills. The topics for group discussion tests are also based on the tacit principle that the content should be either familiar or known and not pose problems in the interaction process. Though the amount of communication (size of contribution) and substantiveness is rated as criteria, content per se is not rated. Group discussion in Level 1 tests lower order interaction skills that are discernible at the conversational level.

The groups discussion test has been modelled on the lines of the Bagrut group discussion test with some modifications (see Shomay, E., Reves, T. and Bejerano, Y. 1986 and Gefen, R. 1987). In Level I the topics are of matters that either concern or pose a problem to the test takers as UKM students. Hence there is sufficient impetus to talk about them and this 'guarantees' initiation by all members of the group in the discussion. Topics in the form of statements are distributed just before the tests from a prepared pool of topics. Each topic comes with a set of questions. Students are allowed to read the questions in advance but discussion on the topic and questions before the test is not permitted. These questions function as cues to direct and manage the interaction. They need not be answered. In fact students may want to speak on other aspects of the topic. An example of the topic and questions is as follows:

Scholarships should be awarded on need and not on merit.

(a) Are both equally important considerations?

(b) Should students have a say in w' o gets scholarships ie. have student representatives on scholarship boards?

(c) Do generous scholarships make students dependent on aid?

(d) Are repayable-upon-graduation loans better than scholarships as more students can benefit?

Groups are small and students are divided (depending on class size) into 4-5 (maximum) students per group. It has been possible to establish a rough ratio between rating time per test-taker and their number per group. Groups of 4 took 15-20 minutes to round off the discussion and groups of 5 took about 20-25 minutes. However, it is desirable not to cut off the discussion after 20-25 minutes, as extra time (usually an extra 5 minutes) helped to confirm ratings. Rating is immediate on the score sheets prepared for the test (see Appendix C ii). A variation of the topics with maximum backwash effect on learning is to use books that have been recomr.,ended for extensive reading as stimulus for group discussion. This has been trialled as a class activity.

It can be seen that the oral interview test is noticeably absent in the sampling of speech interactions for Level I of the course and probably begs the question why, as it is a common and well established test for testing oral interaction. Suffice to say that it is firstly one of the tests administered in the first year integrated program (and therefore sampled). Secondly the group discussion appears to be a more valid (face and content) test of oral interaction in relation to the course objectives.

3 Since a premium is placed on intelligibility/comprehensibility the end-of-semester project tests for overall verbal communicative ability in the rehearsed speech genre in the form of a news magazine that is audio taped for assessment and review. The news magazine may be presented either as a collage of items of news and views of events and activities on campus or thematically eg. sports on campus, cultural activities, student problems etc.

Level II

This level consists of 4 test types.

1 Group discussion

2 Public speaking

3 Debates

4 End-of-semester project

1 In the second level the group discussion test is administered early in the semester and the results used to determine how much more practice is needed in improving interaction skills before proceeding to the more formal performance-oriented speech genres. The topics for the group discussion in the second level are of a more controversial nature than in the first. Although cognitive load is expected to be greater in the tests, procedures for test administration and scoring are the same.

2 Public speaking is tested mid-way in the second semester after lecture-demonstrations and a series of class presentations. As a test of global communication skills, both verbal and non-verbal, it represents fairly high level order skills on the language learning continuum assumed for the course. Like debates, it is a sample of rehearsed speech in a formal situation. It is also viewed as a necessary advanced oral skill. Examples of topics are,

Mothers should not go out to work.
Alcoholism is a worse social evil than drug abuse.

3 The debate is placed at the end of the semester and usually viewed by the students as a finale of sorts of their oral communication skills. As with the public speaking test, topics and teams (for the debates) are made known well in advance and students work on the topics cooperatively for the latter. The backwash effect on the acquisition of social and study skills is tremendous as students are informed that ratings reflect group effort in the debating process. Both tests 2 and 3 are rated immediately and video taped for both review and record purposes.

4 The end-of-semester can take two forms - that of a form of a platform activity (in the public speaking mode) or a chat show (speech interaction). Both test for skills learned or acquired during the course. The platform activity and the formal speech situation can be either an appeal (for blood donation, funds, etc) or the promotion of a product/service or idea. The chat show tests for oral interaction in the form of an extended interview of a 'celebrity'. Both tests simulate real life situations and allow for creativity and flexibility in that students can assume personae.

Criteria and Rating Scales

> "Testers should construct their own rating scales according to the purpose of the test". (Shohamy E. 1988)

Rating scales have been constructed for all the tests developed. A look at the criteria and the rating scales (see appendices) for the various tests discussed above, shows that the criteria for each test varies although some (mainly linguistic) recur as each test samples different types of communicative ability.

Working over a period of time (ie two years = four semesters) it has been possible to specify what criteria should be used to rate each test and therefore what sorts of rating scales to produce. It has also been possible to select specific components from the broader criteria identified for each rating scale. In this sense each test has evolved pedagogically (mainly) and psychologically over a period of time to become more comprehensive in terms of the test (task) specifications. Feedback in the form of student responses (and reaction) to each task has also helped the tests to jell as they were used to make changes especially to the criteria and subsequently the rating scale so as to reflect a wider possible range of responses for each test.

Obviously comprehensiveness of criteria should not be at the expense of the feasibility of rating scales and the practicality of scoring procedures. Too many descriptors can make it difficult for a rater to evaluate the performance in any one task. Using all these simultaneously to make an immediate judgement is no mean task. Hence, instead of fully descriptive qualitative scales, more parsimonious rating scales were devised. Working hand in hand with a checklist of what are essentially holistic criteria which will vary according to test purpose, the tester rates analytically on a 1 to 4 or 6 point scale depending on the test. These scales are also grouped into 3 broad bands of 'weak', 'fair' and 'good' which provide guidelines to help the rater to keep on course in the absence of banded descriptors. There is also space on each score-sheet for tester comments. This allows the tester to make relevant remarks of each test on an individual basis particularly with reference to those factors that had an apparent effect on test performance, verbal, non-verbal or affective.

The problem (personal experience) with banded qualitative rating scales is that the descriptors may not fit the description of the individual student in that some of the performance variables for any one component may be absent while others may be present. And there are students whose performance defy 'pigeon-holing'. However, it is possible to categorize the same students, firstly, on a broad basis as 'weak', 'fair' and 'good' and then work from there to rate them analytically on weighted 6 point scales in this case. It may even be possible to describe them with reference to the criteria on an individual basis as it is small scale testing. While such rating procedures remain subjective and may even be criticized on that basis, at the very least they prevent stereo typing of students by not assigning their performance to prescriptive ready-made bands.

CONCLUSION

Test Anxiety

A certain amount of anxiety has been removed from the testing situations in the course firstly, because of the ongoing nature of the assessments and secondly because of the wider sampling of the speech genres.

> "There is ... evidence in the literature that the format of a task can unduly affect the performance of some candidates. This makes it necessary to include a variety of test formats for assessing each construct... In this case, candidates might be given a better chance of demonstrating potentially differing abilities (Weir, C. 1989).

Practitioners know that not only do levels of test anxiety vary from situation to situation and from testee to testee, it may not even be possible to eliminate anxiety as an affective variable. However, in order to further reduce test anxiety and to 'bias for best', students are informed at the beginning of each level about course objectives and expectations, test types and task specifications explained. Feedback is also provided after each test although actual scores obtained are not divulged.

Other Matters

All tests of courses on the university curriculum (cumulative or otherwise) are seen as achievement tests with scores and grades awarded accordingly. There is a certain amount of tension between rating according to specified criteria and the subsequent conversion of the weightage of the components of these criteria into scores. However despite this constraint it is still possible to speak of a student's profile of performance in the oral communication class from level to level. At the end of the second year similar judgements can be made of them as potential students for the B A in English Studies.

The oral communication course has also been offered more recently as an elective to other students and therefore involves more teachers. While the difference in clientele does change some of the course's methodological perspectives, the objectives have still been maintained as needs are broadly similar. The tests are now being subjected to a process of small-scale teacher validation since the question of some extrapolation is apparent. There have been informal training and practice sessions for the teachers in the use of the criteria and rating scales. Past samples of performance have been reviewed to arrive at bench marks and pre-marking sessions held to increase intra and inter-rater reliability. The intersubjectivity and teacher feedback on all these aspects are invaluable in improving the efficacy of the test as instruments, at least with reference to face and content validity. Obviously more work has to be done before anything conclusive can be said.

REFERENCES

Alderson, C. 1981. Report of the Discussion on Communicative Language Testing in Alderson, C. and Hughes, A. (eds) in ***Issues in Language Testing. ELT Documents III.*** *The British Council.*

Alderson, C. 1983. "Who Needs Jam?" ***Current Developments in Language Testing*** *London: Academic Press. 55-65.*

Bachman, L. (Forthcoming). ***Fundamental Considerations in Language Testing*** *Reading. Mass: Addison-Wesley.*

Bartz, W H. (1979). "Testing Oral Communication in the Foreign Language Classroom" in ***Language in Education: Theory and Practice.*** *Arlington. Virginia: Centre for Applied Linguistics 1-22.*

Brown, G. and Yule, G. 1983. ***Teaching the Spoken Language,*** *Cambridge: Cambridge University Press.*

Callaway, D R. 1980. "Accent and the Evaluation of ESL Oral Proficiency" in Wover, J. Perkins, K. ***Research in Language Testing,*** *Rowley. Mass: Newbury House.*

Carroll, B J. 1980. ***Testing Communicative Performance,*** *Oxford: Pergamon Press.*

Clark, J L D. 1975. "Theoretical and Technical Considerations in Oral Proficiency Testing in Jone, R. and Spolsky, B. (eds) ***Testing Language Proficiency*** *Arlington. Virginia: Centre for Applied Linguistic 10-28.*

Gefen, R. 1987. "Oral Testing --- The Battery Approach" ***Forum*** *25.2 24-27*

Kreshen, S. and Terrell T. 1983. ***The Natural Approach: Language Acquisition in the Classroom,*** *Pergamon: Oxford.*

Nair-Venugopal, S. 1988. "Simulations, authentic oral language use in multiracial settings" ***Babel Occasional Papers*** *1 27-31*

Skehan, P. 1988. Language Testing: Part I 1-10. Language Testing: Part II 1-13. ***Language Teaching Abstracts.***

Shohamy, E. 1988. "A Proposed Framework for Testing the Oral Language of Second/Foreign Language Learners" ***Studies in Second Language Acquisition*** *10.2 165-180*

Shohamy, E., Reves. T. and Bejerano, Y. 1986. "Large Scale Communicative Language Testing: A Case Study" in Lee, Fok, Lord and Low (eds) ***New Directions in Language Testing*** *Oxford: Pergamon Press.*

Spolsky, B. 1975. "Language Testing: Art or Science" Lecture Delivered AILA ***World Congress*** *'75 University of Stuttgart.*

Weir, C. 1988. "Approaches to Language Test Design: A Critical Review" Lecture Delivered at British Council Specialist Course 937 ***Testing Oral Interaction: Principles and Practice,*** *Reading, England.*

WHAT WE CAN DO WITH COMPUTERIZED ADAPTIVE TESTING... AND WHAT WE CANNOT DO!

Michel Laurier

INTRODUCTION

Among numerous applications of computers for language teaching and learning there is a growing interest for a new acronym: CAT which stands for Computerized Adaptive Testing. CAT can be seen as the second generation of computerized tests (Bunderson, Inouye & Olsen 1989). The first generation consisted of conventional test administered by computers; further generations will be less obtrusive and will provide constant advice to the learners and teachers. In this paper we shall attempt to explain how CAT works and what is the underlying theory. The various steps involved in implementing an adaptive test will be described with examples from a placement test that we have developed in French.

PRINCIPLES OF ADAPTIVE TESTING

Computers in testing are particularly useful because of two advantages over conventional testing methods:

- number-crunching capabilities: Conventional marking systems often means counting the number of right answers or converting a score with a pre-set scale. Using a computer allows more complex assessment procedures right after the test or even during the test. These calculations may use the data that is available more efficiently. In addition, computers are fast and virtually error-free.

- multiple-branching capabilities: Using "intelligent" testing systems, some decisions can be made during the administration of the test. The computer can analyze students' responses and decide which item will be submitted, accordingly. Therefore, the inherent linearity of a conventional test is no longer a limitation.

CAT takes full advantage of these two properties of the computer. Let's suppose we want to assign a student to a group that would suit his needs by means of a conventional placement test.

We do not know a priori at which level the student could be placed; he/she could be an absolute beginner in the language or an "educated native". In this case, the test should probably include some difficult items, as well as some easy ones. In fact, given the student's level, how many of the items of a two hour test are relevant? Probably less than 25%. Some of the items will be too easy, particularly if the student is at an advanced level. From the student's point of view, those items are boring, unchallenging; from the psychometric point of view, they do not bring valuable information because the outcome is too predictable. On the other hand, some items will be too difficult, particularly for beginners who will feel frustrated because they find that the test is "over their heads"; again, there is very little information on the student's level that can be drawn from these items.

Adaptive testing has also been called "tailored testing" because it aims at presenting items that suit the students' competence and that are informative. In an open-ended test, this means items in which the chance to answer correctly will be approximately fifty/fifty. This approach to testing problems might bring to mind Binet's multi-stage intelligence tests that were developed at the beginning of the century. For language teachers, it may also resemble recent oral interview procedures in which the examiner is encouraged to adapt the exchange to the examinees' performance (Educational Testing Service 1985).

Adjusting the test is in fact a complex process that CAT seeks to replicate. For this task, we need:

- an item bank: a collection of items stored with some specifications and measuring the same ability at different levels.

- a selection procedure: an algorithm which will choose and retrieve the most appropriate item at a given moment, with a given examinee.

ITEM RESPONSE THEORY

Although different theoretical frameworks could be applied to set up the item bank and the selection procedure, the most widely used is the Item Response Theory (IRT). Despite its mathematical complexity, IRT is conceptually attractive and very interesting for CAT. The theory was first labeled "latent trait theory" by Birnbaum (1968) because it assumes that a test labeled "latent trait theory" by Birnbaum (1968) because it assumes that a test score or a pattern of answers reflects a single construct that is not directly observable. What the test measures is known as the "trait" and corresponds to the subject's ability. The theory was refined by F. Lord who studied the "Item Characteristic Curve" (Lord 1977). "An item characteristic curve (ICC) is a mathematical function that relates the probability of success on an item to the ability measured by the item set or test that contains it" (Hambleton and Swaminathan 1985:25). If we plot the probability of answering correctly against the examinees' ability, the curve should rise as the ability level increases. Thus, the probability of having a right answer at the advanced level will be very high but should be very low at the beginner's level. The ability is expressed in terms of standard deviations and ranges from roughly -3 to +3. Figure 1 shows the curve for an "Intermediate" level item. The inflection point of this ICC is around 0 which corresponds to the sample mean. Since the subject's ability and the item difficulty are expressed on the same scale, we say that the difficulty of the item (the parameter b) is 0.

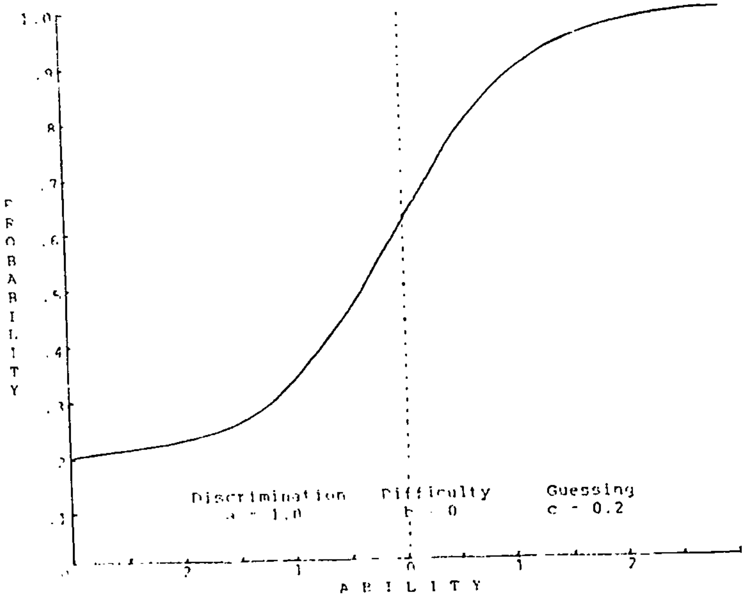


Figure 1: Item Characteristic Curve

If an item clearly separates the advanced students from the beginners the curve should be very steep; if it does not, the curve will be flatter. In other words, the slope of the ICC corresponds to the discrimination (the parameter a). An item with a discrimination index of 1 or more is a very good item. Finally, we see that, in this particular case, the curve will never reach the bottom line. This is due to the fact that the item is a multiple choice question which involves some guessing.

This is expressed with a third parameter (parameter c). A m/c item with five options should have a c around .2. Of course, in reality, such a regular curve is never found. The degree to which the data for an item conforms to an ICC is the "item fit". Misfitting items should be rejected.

Once the parameters are known, we can precisely draw the ICC using the basic IRT formula

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

where 0(theta) represents the subject's ability and D a scaling constant set at 1.7. A simpler formula for a less complex but generally less accurate model has been proposed by G. Rasch (1960). The Rasch model is a one-parameter model; it assumes that there is no guessing and that all the items discriminate equally. Under this model, only the difficulty has to be estimated.

The parameter estimation is a complex mathematical procedure that requires a computer. Various programs are available either on mainframe computers (v.g LOGIST, Wingersky, Barton & Lord 1982) or micro-computers (v.g. MicroCAT, Assessment Systems Corp. 1984). To estimate the parameters properly, particularly with the three-parameter model (discrimination, difficulty and guessing) a large sample is needed - about 1,000 examinees. Fortunately, the distribution of the sample does not have to reflect exactly the distribution of the population because the program will try to fit a curve rather than calculate proportions of correct answers. The item calibration is sample-free. This property of IRT models is known as the "invariance of items". IRT provides also the "invariance of subjects" which means that we get test-free person measurement. This second property is crucial in adaptive testing because it implies that ability estimates can be calculated and compared even though different items have been submitted.

IMPLEMENTATION OF THE TEST

The following steps are involved in creating the item bank:

- Planning the bank: Are we measuring more than one common trait? If so, then several item banks should be set up. At this stage, we must also make sure that the items can be administered, answered and marked both with a "paper-and-pencil" format and with a computerized version. Since field testing is expensive, a great deal of attention must be paid to the wording of the items. For large item banks, several versions using "anchor items" will be necessary.

- Field testing and item analysis: The items will be tried out on a small sample - 100 to 200 subjects. Classical item analysis using proportions of correct answers and correlations is helpful in order to eliminate bad items from the next version. At this stage, some dimensionality analysis can be conducted to make sure the test (or sub-test) is measuring a single trait.

- Field testing and calibration: The new version(s) is(are) administered to a large sample - 200 to 2,000 depending on the model chosen and the quality of the sample. This data will be processed so that item parameters and degree of fit will be obtained for each item.

- Inclusion to the bank: If the item is acceptable, it will be added to the bank. At least, an identification code, the questions (and the options with multiple-choice items), the right answer and the parameters should appear on an item record. Additional information may be incorporated (Henning 1986).

Of course, a management system will have been previously set up. A management system works like a data base system. Each sub-test is a data base that can be accessed with the management system. Once the user has chosen a sub-test, different operations can be executed:

- Updating the bank: new items may be added, some others deleted. The user should also be able to browse in the bank and modify an item without having to rewrite it.

- Importing items: When a set of items are located in another file, there should be provisions to execute a mass transfer into the bank.

- Listing the items: Each item can been seen individually on the screen. Yet the user can also call a list of the items. Each line will show the identification code of an item, the parameters, and a cue to remind the question.
  In addition, our system calculates the "Match index". According to Lord (1970), this value corresponds to the ability at which the item is the most efficient.

- Obtaining the item information: Under IRT, one can tell how much information can be obtained at different points of the ability scale. As the information sums up, at a specific ability point, the estimation becomes increasingly more reliable at this point.

The selection procedure is a method that can be applied in order to estimate the examinee's ability after an answer and to find the next item that is the most appropriate. The concept of item information is crucial as the most appropriate item is the one that brings the most information for a given ability. Tracing the administration of the adaptive test we have designed will help to understand how the program works. We needed a computerized placement test for English speaking post-secondary students learning French as a second/foreign language in Canada. As a placement test, the instrument attempts to assess the student's general proficiency. It assumes that such a construct exists even though a more refined evaluation should probably divide this general competence in various components such as the grammatical competence, the discourse competence or the sociolinguistic competence (Canale and Swain 1980). The format of the test is affected by the medium, the micro-computer. The three sub-tests contain multiple-choice items because we want to minimize the use of the keyboard and because open-ended answers are too unpredictable to be properly processed in this type of test. The organization and the content of the test also reflect the fact that we had to comply with IRT requirements.

THE ADMINISTRATION OF THE TEST

Within the IRT framework, procedures have been developed to estimate the student's ability, using the answers to the items and the parameters of these items. However, calculating the student's ability is not possible when the program is started since no data is available. This is the reason why, at the beginning of the test, the student will be asked some information about his/her background in the second/foreign language:

> How many years did the student study the language?
> Did he/she ever live in an environment where this language is spoken?
> If so, how long ago?

Then the program prompts the student to rate his/her general proficiency level on a seven category scale ranging from "Beginner" to "Very advanced". All this information is used to obtain a preliminary estimation that will be used for the selection of the first item of the first sub-test. Tung (1986) has shown that the more precise is the preliminary estimation, the more efficient is the adaptive test.

The first sub-test contains short paragraphs followed by a m/c question to measure the student's comprehension. According to Jafarpur (1987), this "short context technique" is a good way to measure the general proficiency. Figure 2 illustrates how the adaptive procedure works. At the beginning of the sub-test, after an example and an explanation, an item with a difficulty index close to the preliminary estimation is submitted.

Item	U	a	b	c	Score	Theta	Info.	Error
C021	0	1.212	0.702	0.240	0/1	-0.750	?	?
C027	0	0.982	-0.819	0.231	0/2	-0.950	?	?
C041	1	0.909	-0.930	0.264	1/3	-1.833	0.338	1.719
C037	1	1.346	1.109	0.219	2/4	-1.129	1.948	0.716
C032	1	0.967	-1.109	0.180	3/5	-0.894	2.685	0.610
C022	0	1.005	-0.568	0.250	3/6	-1.070	2.752	0.603
C034	1	0.807	0.905	0.228	4/6	-0.946	3.269	0.553
C030	0	1.220	0.809	0.198	4/7	-1.148	3.408	0.542

Figure 2 - Items used in sub-test #1

In the example, the first item was failed (U = 0) and the program then selected an easier one. When at least one right and one wrong answer have been obtained, the program uses a more refined procedure to calculate the student's ability. The next item will be one which has not been presented as yet and that is the closest to the new estimation. The procedure goes on until the pre-set threshold of information is reached. When this quantity of information is attained, the measure is precise enough and the program switches to the next sub-test.

The same procedure is used for the second part with the estimation from the previous sub-test as a starting value. On the second sub-test, a situation is presented in English and followed by four grammatically correct statements in French. The student must select the one that is the most appropriate from a semantic and sociolinguistic point of view. Raffaldini (1988) found this type of situational test a valuable addition to a measure of the proficiency. Once we have obtained sufficient information, the program switches to the third sub-test, which is a traditional fill-the-gap exercise. This format is found on most of the current standardized tests and is a reliable measure of lexical and grammatical aspects of the language. Immediately after the last sub-test, the result will appear on the screen.

Since a normal curve deviate is meaningless for a student, the result will be expressed as one of the fourteen labels or strata that the test recognizes along the ability range: "Absolute beginner, Absolute beginner +, Almost beginner ... Very advanced +".

ADVANTAGES AND LIMITATIONS

Both the students and the program administrators appreciate that the result is given right away. The students receive immediate feedback on what he/she did and the result can be kept confidential. Since there are no markers, the marking is economical, error-free and there is no delay. Individual administration as opposed to group administration is, in some situations, an asset: the students can write the test whenever they want, without supervision. Because of the adaptive procedure, the tests are shorter. In order to reach a comparable reliability with our test, we need a "paper-and-pencil" version that is at least twice as long as the CAT one. Actually, in most cases, CAT will use only 40% of the items of the equivalent conventional test. Finally, the adaptive procedure means that the student is constantly faced with a realistic challenge: the items are never too difficult or too easy. This means less frustration, particularly with beginners. With a more sophisticated instrument than the one we designed, one could even find other positive aspects of CAT. For example, with IRT it is possible to recognize misfitting subjects or inappropriate patterns and therefore detect phoney examinees. Taking advantage of the capabilities of the computer, one could also make the testing environment more enjoyable.

However, there are also very serious limitations with CAT. Even with the fanciest gadgetry, the computer environment will always be a very artificial one. It is always a remote representation of the real world and precludes any form of direct testing. Moreover, the type of answer is restricted because of the machine itself and because of the psychometric model. With the combination of the present technology and IRT, it is hard to imagine how a test could use anything other than m/c items or very predictable questions. The medium, the computer, not only affects the type of answers but also the content of the test. In our test, we wanted to use standard and affordable hardware but some students complained that the test was very poor in assessing oral skills. In spite of recent innovations with videodiscs, audio-tape interfaces, CD-Rom, or even artificial speech devices, the stimulus in CAT is generally written. On the other hand, the model, IRT, not only affects the type of answer but also the practicality of the development. In our test, three parts of fifty items each were administered to more than 700 hundred examinees. This is considered as a minimum and some research shows that even with 2,000 examinees, the error component of a three-parameter calibration may be too large. Using a Rasch model may help to reduce the sample size, usually at the expense of the model fit, but the field testing will always be very demanding. Therefore, CAT is certainly not applicable to small-scale testing.

Perhaps the most formidable problem, is the assumption of unidimensionality. This concept refers to the number of traits that are measured. Under IRT, a common dimension, ie. a single factor, must clearly emerge. Otherwise, applications of IRT may be highly questionable. Even though the calibration procedure is statistically quite robust and most language tests will comply with the unidimensionality requirement (Henning, Hudson & Turner 1985), many testing situations are based on a multidimensional approach of the language competence (Bachman, forthcoming).

Multi-dimensional calibration techniques exist but they are not always practical (Dandonelli & Rumizen 1989). One particular type of unidimensionality is the independence of the items. This principle implies that an answer to one item should never affect the probability of getting a right answer on another item. Cloze tests usually do not meet this requirement because finding a correct word in a context increases the chance of finding the next word.

Finally, when all the theoretical problems have been solved some practical problems may arise. For example, for many institutions the cost of the development and implementation of an adaptive test could be too high. Madsen

(1986) studied the student's attitude and anxiety toward a computerized test; attention must be paid to these affective effects.

CONCLUSION

These limitations clearly indicate that CAT is not a panacea. It should never be used to create a diagnostic test that aims at finding weaknesses or strengths on various discrete points because this type of test is not unidimensional. By the same token, it should not be used on so-called "communicative" tests that attempt to measure aspects of the communicative competence without isolating the different dimensions in separate sub-tests. Canale (1986) mentions that the testing environment is so artificial that CAT lacks validity when test results are used to make important decision - for a certification test, for instance.

However if only a rough estimation over a wide range of ability is needed, for placement purposes, for example, CAT may be a very adequate solution. It is also appropriate if the trait being measured is unique such as general proficiency, vocabulary, grammar. It could also be a solution to testing problems for some integrative tests of receptive skills particularly if the result will not affect the student's future or can be complemented with more direct measures.

In short, a CAT will always be a CAT, it will never be a watchdog.

NOTES

[1] For an excellent introduction to IRT, see Baker (1985)

[2] An experimental version of this test has been developed at the Ontario Institute of Studies in Education (Toronto) and will be implemented at Carleton University (Ottawa).

REFERENCES

Assessment Systems Corporation (1984) User's Manual for the MicroCAT testing system. St.Paul, MN.

Bachman L.F. (forthcoming) Fundamental considerations in language testing. Reading, MA: Addison-Wesley.

Baker F.B. (1985) The basics of item respons. theory. Portsmouth, NH.

Birnbaum A. (1968) Some latent trait models and their use in infering an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bunderson C.V., Inouye D.K. & Olsen J.B. (1989) The four generations of computerized educational measurement. In R.L. Linn(Ed.) Educational Measurement 3rd ed. (pp. 367-408) New York: American Council on Education - Macmillan Publishing.

Canale M. (1986) The promise and threat of computerized adaptive assessment of reading comprehension. In C. Stansfield (Ed.) Technology and language testing (pp. 29-46). Washington, D.C.: TESOL.

Canale M. & Swain M. (1980) Theoritical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1, 1-47.

Dandonelli P. & M. Rumizen (1989) There's more than one way to skin a CAT: Development of a computer-adaptive French test in reading. Paper presented at the CALICO Conference, Colorado Spring, CO.

Educational Testing Service (1985) The ETS Oral Interview Book. Princeton, NJ.

Hambleton and Swaminathan (1985) Item response theory: Principles and applications. Boston, MA: Kluwer Academic Publishers.

Henning G. (1986) Item banking via DBase II: the UCLA ESL proficiency examination experience. In C. Stansfield (Ed.) Technology and language testing (pp. 69-78). Washington, D.C.: TESOL.

Henning G., Hudson T. & Turner J. (1985) Item response theory and the assumption of unidimensionality for language tests. Language Testing, 2, 141-154.

Jafarpur A. (1987) The short-context technique: an alternative for testing reading comprehension. Language Testing, 4, 133-147.

Lord F.M. (1970) Some test theory for tailored testing. In W.H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance (pp. 139-183) New York: Harper & Row.*

Lord F.M. (1977) Practical application of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

Madsen H. (1986) Evaluating a computer adaptive ESL placement test. CALICO Journal, December, 41-50.

Raffaldini T. (1988) The use of situation tests as measure of communicative ability. Studies in Second Language Acquisition, 10, 197-215.

Rasch G. (1960) Probabilistics models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.

Tung P. (1986) Computerized adaptive testing: implications for language test developers. In C. Stansfield (Ed.) Technology and language testing (pp. 11-28). Washington, D.C.: TESOL.

Wingersky M.S., Barton M.A. & Lord F.M. (1982) LOGIST user's guide. Princeton, NJ: Educational Testing Service.

LIST OF CONTRIBUTORS

J Charles Alderson
Department of Linguistics &
 Modern Languages
University of Lancaster
Bailrigg, Lancaster LA1 4YW
United Kingdom

Geoff Brindley
National Centre for English
 Language Teaching & Research
Macquarie University
North Ryde, NSW 2109
Australia

James Dean Brown
H Gary Cook
Charles Lockhart
Teresita Ramos
Department of ESL
University of Hawaii at Manoa
1890 East West Road
Honululu, HI 96822
USA

Peter Doye
Technische Universitat Braunschweig
 Seminar fur Englishe und
 Franzosische Sprache
Buitenweg 74/75
3300 Braschweig
Germany

David E Ingram
Centre for Applied
 Linguistics & Languages
Griffith University
Nathan, Queensland 4111
Australia

Michel Laurier
National Language Testing Project
Department of French
Carleton University
Ottawa, Ontario K1S 5B6
Canada

Liz Hamp-Lyons
Sheila B Prochow
English Language Institute
3001 North University Building
University of Michigan
Ann Arbor, Michigan 48109-1057
USA

T F McNamara
Language Testing Unit
Department of Linguistics
& Language Studies
University of Melbourne
Parkville, Victoria 3052
Australia

Michael Milanovic
Evaluation Unit
University of Cambridge
Local Examinations Syndicate
Syndicate Buildings
1 Hills Road
Cambridge CB1 2EU
United Kingdom

Keith Morrow
Bell Education Trust
Hillscross, Red Cross Lane
Cambridge CB2 2QX
United Kingdom

John W Oller Jr
Department of Linguistics
University of New Mexico
Albuquerque, NM 87131
USA

Don Porter
Centre for Applied Linguistics
University of Reading
Whiteknights, Reading RG6 2AA
United Kingdom

John Read
English Language Institute
Victoria University of Wellington
P O Box 600, Wellington
New Zealand

Charles W Stansfield
Division of Foreign Language
 Education & Testing
Center for Applied Linguistics
1118 22nd Street, N W
Washington DC 20037
USA

Shanta Nair-Venugopal
English Department
Language Centre
Universiti Kebangsaan Malaysia
Bangi, Selangor, Darul Ershan 43600
Malaysia