ELT documents

111- Issues in Language Testing

The British Council

ELT documents
111- Issues in Language Testing

Editors: J Charles Alderson
Arthur Hughes

The British Council
Central Information Service
English Language and Literature Division

ELT Documents is now including a correspondence section. Comments arising from articles in current issues will therefore be most welcome. Please address comments to ELSD, The British Council, 10 Spring Gardens, London SW1A 2BN.

ISBN 0 901618 51 9

CONTENTS

SECTION 3: **General Language Proficiency**

INTRODUCTION

This book arose from an occasion in October 1980 when seven applied linguists met in Lancaster to discuss what they felt were important problems in the assessment of learning a second or foreign language. This Symposium resulted, partly because of its informal nature and its deliberately small size, in an intense discussion in certain areas, a concentration which is rarely possible in conferences or large seminars. It was felt that the Symposium had been so useful that it was decided to make the discussion public, in order not only to let others know what had happened at Lancaster, but also to encourage and stimulate a much broader and hopefully even richer debate in the areas touched upon.

Testing has become an area of increased interest to language teachers and applied linguists in the last decade. Yet as Davies says (Davies 1979) testing has for many years firmly resisted attempts to bring it within the mainstream of applied linguistics. This is no doubt to some extent due to historical reasons, as both Davies and Morrow (this volume) suggest. In the era that Spolsky dubbed the 'psychometric-structuralist period' language testing was dominated by criteria for the establishment of educational measuring instruments developed within the tradition of psychometrics. As a result of this emphasis on the statistical analysis of language tests, a group developed, over the years, of specialists in language testing, 'Testing Experts', popularly believed to live in an arcane world of numbers and formulae. As most language teachers are from a non-numerate background (sometimes having deliberately fled 'figures') it is not surprising that they were reluctant to involve themselves in the mysteries of statistics. Consequently, an expertise developed in language testing and particularly proficiency testing, divorced from the concerns of the language classroom, and imbued with its own separate concerns and values which to outsiders were only partially comprehensible and apparently irrelevant. Despite the advent of Spolsky's third phase of language testing — the psycholinguistic-sociolinguistic phase (what Moller (this volume) calls the third and fourth phases — psycholinguistic-sociolinguistic and sociolinguistic-communicative phases) — 'testing' has not yet recovered from this image of being stubbornly irrelevant to or unconcerned with the language teacher, except for its embodiment in 'exams' which dominate many a syllabus (be it the Cambridge First Certificate or the TOEFL). Teachers who **have** felt they should be concerned with assessing what or whether learners have learned have found the jargon and argumentation of 'Testing' forbidding and obscure.

But evaluation (note how the terminology has changed over the years, with the intention of making the subject less threatening) is readily acknowledged by teachers and curriculum theorists alike to be an essential part of language learning, just as feedback is recognised as essential in any learning process. The consequence of this need to evaluate has been the fact that teachers have actually carried out tests all along but have felt uncomfortable, indeed guilty and apologetic about doing so when there is apparently so much about 'testing' they do not know. So when suggesting that 'Testing' has become more central to the present-day concerns of language teachers, it is not intended to imply that previously – 'in the bad old days' – nobody tested, or that the testing that was done was of ill repute, but merely to suggest that teachers felt that what they were doing was in some important sense lacking in respectability however relevant or important it might actually have been. The fact is, however, that testing has become an area of increased research activity, and many more articles are published on the subject today in professional journals than ten years ago. This is evidence of a turning in the tide of applied linguistics towards more empirical concerns.

It has been suggested that testing has to date remained outside the mainstream of applied linguistics; in particular, the view of language incorporated in many tests has become increasingly at odds with theories of language and language use – indeed, to some extent at least, it no longer reflects classroom practice in language teaching. Now there may be good arguments for tests not to follow the whim of fashion in language teaching, but when there is a serious discrepancy between the teaching and the means of evaluating that teaching, then something appears to be amiss. The feeling abroad today is that theories abound of communicative language teaching, of the teaching of ESP, of integrated language teaching, but where are the tests to operationalise those theories? Where are the communicative language tests, the ESP tests, the integrated language tests? Applied linguists and language teachers alike are making increasingly insistent demands on language testers to supply the language tests that current theory and practice require, and the response of testers has, to date, been mixed. Some have rushed in where others have feared to tread: extravagant claims have been made for new techniques, new tests, new assessment procedures. Others have stubbornly resisted the pressure, claiming that tests of communicative competence or ESP are either impossible (in theory, or in practice) or unnecessary because existing tests and techniques are entirely adequate. Inevitably, there are also agnostics on the side lines, who remain sceptical until they have seen the evidence for and against the claims of either side.

This book is for those agnostics, though believers and non-believers alike may find something of interest. The Symposium at Lancaster was an attempt to focus, without taking sides, on areas of major concern to teachers and testers at present:

communicative language testing,
the testing of English for Specific Purposes,
the testing of general language proficiency.

It was hoped by intense debate to establish what the important issues were in these areas, so that the interested reader could provide himself with a set of criteria for judging (or constructing) language tests, or perhaps more realistically, for investigating further. It is clear, always, that more research is needed but it is hoped that this book will help to clarify where research and development needs to be concentrated at present. We are living in a world of claim and counter-claim, where the excitement of the battle may make us lose sight of the reasons for the conflict: namely the need for learners and outsiders to assess progress in language learning or **potential** for such progress, as accurately as possible. No research programme or test development should forget this.

The format of the Symposium was as follows. Having decided on the three main areas for debate, recent and influential articles in those areas were selected for study and all Symposium participants were asked to produce papers reacting to one or more of these articles, outlining what they felt to be the important issues being raised. These reaction papers were circulated in advance of the Symposium, and the Symposium itself consisted of a discussion in each of the three areas, based on the original articles and the related reaction papers.

Like the Symposium, the volume is divided into three main sections: one section for each of the areas of communicative language testing, ESP testing, and general language proficiency. Within each section there are three parts: the original article(s), the reaction papers and an account of the discussion based upon tape recordings of the proceedings by the present writer. These accounts of the discussion do not represent the views of any one participant, including the present writer, but are an attempt to summarise the issues that were raised. However, it should be stressed that although the accounts of the discussion attempt to be fair to the substance and quality of the debate, they must, inevitably, ultimately represent one person's view of what was said, since it would be impossible to achieve complete consensus on what was said, let alone its correctness or significance. At times the accounts repeat points made in the reaction papers also published in this volume, but no apologies are offered for repetition, as this simply reflects the level of interest in or

concern over these particular points. Although it was hoped to include responses from the authors of the original articles only one response was available at the time of going to press, that of Helmut Vollmer. Nevertheless, it is hoped that subsequent debate will include the responses and further thoughts of the other authors in the light of these discussions.

This is not a definitive volume on language testing – and it does not attempt to be such. What this book hopes to do is to encourage further debate, a critical or sceptical approach to claims made about 'progress' and 'theories', and to encourage practical research in important areas.

It has not been the intention of this Introduction to guide the reader through the discussions – that would have been presumptuous and unnecessary – but rather to set the scene for them. Thus there is here no summary of positions taken, arguments developed and issues raised. However, there is, after the three main sections, an Epilogue, and the reader is advised not to ignore this: it is intended, not to tell the reader what he has read, but to point the way forward in the ongoing debate about the assessment of language learning. 'Testing' should not and cannot be left to 'Testers': one of the most encouraging developments of the last decade is the involvement of more applied linguists in the area of assessment and evaluation. In a sense, there can be no Epilogue, because the debate is unfinished, and we hope that participation in the debate will grow. It is ultimately up to the reader to write his own 'Way Forward'.

Thanks are due to all Symposium participants, not only for their contributions, written and spoken, to the Symposium, but also for their help in preparing this volume. Thanks are also due to the Institute for English Language Education, Lancaster, for hosting the Symposium and contributing materially to the preparation of this book.

J Charles Alderson,
University of Lancaster

SECTION 1

COMMUNICATIVE LANGUAGE TESTING: REVOLUTION OR EVOLUTION?[1]

Keith Morrow, Bell School of Languages, Norwich

Introduction

Wilkins (1976) concludes with the observation that, 'we do not know how to establish the communicative proficiency of the learner' and expresses the hope that, 'while some people are experimenting with the notional syllabus as such, others should be attempting to develop the new testing techniques that should, ideally, accompany it' (*loc cit*). In the two years that have passed since the publication of this book, the author's hope on the one hand has been increasingly realised, and if his observation on the other is still valid, there are grounds for believing that it will not be so for much longer.

At the time of writing, it is probably true to say that there exists a considerable imbalance between the resources available to language teachers (at least in E F L) in terms of teaching materials, and those available in terms of testing and evaluation instruments. The former have not been slow to incorporate insights into syllabus design, and increasingly methodology, deriving from a view of language as communication; the latter still reflect, on the whole, ideas about language and how it should be tested which fail to take account of these recent developments in any systematic way.[2]

This situation does seem to be changing, however. A number of institutions and organisations have set up working parties to assess the feasibility of tests based on communicative criteria, and in some cases these have moved on to

---

[1] This article was first published in *The Communicative approach to language teaching* ed: C J Brumfit and K Johnson. Oxford University Press, 1979. Reprinted here by kind permission of Oxford University Press.

[2] Exceptions to this are the two oral examinations promoted by the Association of Recognised English Language Schools: The ARELS Certificate and the ARELS Diploma, as well as the Joint Matriculation Board's Test in English for Overseas Students. But without disrespect to these, I would claim that they do not meet in a rigorous way some of the criteria established later in this paper.

EPILOGUE

Arthur Hughes, University of Reading

The symposium, the structure of which is mirrored by this volume, dealt in turn with three closely related topics. As a result, the same or very similar issues tended to recur, not always in quite the same form, often without their interconnectedness being made explicit. The purpose of the epilogue is to provide a brief summary of these issues, to show how they relate to each other, and to suggest what part they may play in the future development of language testing. In order to do this, instead of treating separately each of the three topics, I shall base what I have to say on the criteria against which all tests, however novel or exciting, must be judged. These are, of course, validity, reliability, and practicality.

As Carroll himself says, the superiority of ELTS over the current test needs to be demonstrated. The ultimate criterion for a test like ELTS is that of **predictive validity.** Its superiority – if indeed it is superior – must be shown in terms of its ability to predict whether an applicant will be able to cope with the linguistic demands made on him by a particular course of study. The problems associated with such validation were discussed at the symposium. But whatever the difficulties, and however persuasive the arguments for giving the test the structure it has, its predictive power has to be demonstrated empirically. It would be particularly interesting to know if, for example, in predicting academic outcomes for science students, an ELTS with a second phase relating to science subjects would prove more accurate than one with a second phase relating to the social sciences. If it did, this would provide powerful support for ESP testing. Until the results of ELTS validation tests are known, however, we must suspend judgement. The ELTS test has a secondary, diagnostic function: to determine the nature and duration of the course of language instruction needed to achieve the required competence in the language. This function, too, is a predictive one and susceptible to similar empirical validation.

By contrast, the RSA test, with which Morrow is associated, to the best of my knowledge makes no claims to prediction. That test and other similar, 'communicative' tests must therefore be subjected to **concurrent validation.** Since there appear to be no comparable tests already validated, this must be based on something like the comparison of scores made on the test by a subset of candidates with ratings of their performance in an extended series of communicative tasks. Once more it has to be said that it is not rhetoric but only empirical validation studies which will convince us of the efficacy of new tests or testing methods.

The proposals that Carroll and Morrow make, and the arguments that they offer in their support, are concerned essentially with **content validity.** Morrow wants communicative language tests to sample the skills involved in communication, while Carroll intends the second part of the ELTS test to sample the tasks that will be required of students on various academic courses, with particular attention being paid to relevant varieties of English. As the symposium discussion revealed, representative sampling of these skills may be very difficult to achieve. For one thing, we lack thoroughly researched and sufficiently detailed analyses of students' language needs on whose inventories sampling might be based.[1] For another, despite Carroll's faith in Munby, we do not have a descriptive framework of language use comparable in completeness or systematicity to those we have of language form; nor do we have anything like a full understanding of the relationships holding between even those elements and dimensions of language use with which we are reasonably familiar. If, however, Carroll and Morrow are successful in their sampling – if, that is, they can predict from the sample of responses obtained in the test to the population of responses in which they are interested – then not only will their tests have greater predictive or concurrent validity (other things being equal), they should also have a beneficial effect on language teaching.

The lack of a demonstrably valid conceptual system on which to base tests of language use, referred to above, may be remedied, at least in part, by **construct validation** studies. If we follow Cronbach (1960) rather than Davies (1968), construct validation is seen as the empirical validation of an interpretation (expressed in terms of underlying concepts) of performance on a test. As such, it may not have immediate appeal to those who regard themselves as 'practical testers' rather than 'testing theoreticians'. Nevertheless, the results of construct validation studies may have important implications for test construction. The better we understand just what underlies performance on language tests, the more confidently we can build new, appropriate tests for particular purposes. The recent upsurge of interest in construct validation owes much to Oller and his promulgation of the Unitary Competence Hypothesis. Verification of this hypothesis would seem to undermine the positions taken in their papers by Carroll and Morrow. In fact, even though research carried out by Oller and his associates has tended to support the hypothesis, there has been criticism (some of it in this volume)

---

[1] Weir is at present working on a large scale study of the language needs of overseas students on behalf of the Associated Examining Board. It remains to be seen what part his analysis will play in the construction of a proposed new examination for such students.

of the methodology and materials used in these studies, as well as of the interpretation of their results. Palmer and Bachman's paper presents counter-evidence; and since the symposium Hughes and Woods (1981) have found as many as four statistically significant components underlying performance on the Cambridge Proficiency Examination. In stimulating so much research, however, Oller has made an invaluable contribution to language testing. It is to be hoped that the current enthusiasm for construct validation studies continues. A great many candidates for investigation, such as **enabling skills** and **language varieties**, have presented themselves in this volume.

A word ought to be said about **face validity.** While face validity is sometimes dismissed as not 'real' validity, it is of undoubted importance in test design. A test's lack of face validity will have a detrimental effect on predictive or concurrent validity; at least some candidates will fail to take the test seriously, and so their performance on the test will not provide an accurate picture of their ability.

There ought to be no argument about the need for test **reliability.** Measurement cannot be consistently accurate if it is not reliable. It may be easier to achieve reliability through the use of a great many discrete items and of techniques which permit objective scoring. But we know that through careful sampling, marking scales based on well defined and recognisable levels of ability, and multiple assessment, it is possible to obtain high reliability for essay questions and interviews. It seems unlikely that problems arising from the techniques to be used in the new communicative tests will not be amenable to similar solutions. The reliability coefficients of these tests will tell us how successful their constructors have been in finding them.

The final criterion is that of **practicality.** Tests cost money to construct, administer, score and interpret. ESP testing implies more tests – and so greater costs – than general proficiency testing; and the achievement of reliability in the assessment of language production skills will be more expensive than is the case with multiple-choice tests. At the same time it has to be recognised that valid tests may save money. If the ELTS test proves successful, one administration might easily save several thousand pounds (and avoid the waste of a year or more of the applicant's life!). We must realise that the weighing of such savings against costs incurred may be as influential in the development and use of new tests as the skills and knowledge of test constructors.

Throughout the symposium, as in this epilogue, there were repeated calls for more research. In their excitement the participants even promised to do some of it themselves. It is only through continued research (in the broadest sense) that the current level of interest in language testing will be

maintained or even increased. It is in recognition of this that a BAAL (British Association for Applied Linguistics) seminar on language testing research, with sections parallel to those of the symposium, is to be held at Reading University in December, 1981. And it is to promote the rapid dissemination of ideas and information on tests and testing research, and to encourage co-operation between researchers, that a Language Testing Newsletter has been established.[2]

As Alderson says in his introduction to this volume, a very great many people have a contribution to make to the future development of language testing. It is hoped that the reader will recognise himself as one of their number.

BIBLIOGRAPHY

CRONBACH, L J
Essentials of Psychological Testing (second edition) New York, Harper and Brothers, 1960.

DAVIES, A, *ed*
Language Testing Symposium. Oxford University Press, 1968.

HUGHES, A and WOODS, A J
Unitary Competence and Cambridge Proficiency Paper presented at AILA Congress, Lund, August, 1981.

[2] Information on the Newsletter can be had from the writer, at the Department of Linguistic Science, University of Reading, Reading, RG6 2AA.

contents

ISBN 0 08 030301 3