Studies in Language Testing 18

European Language Testing in a global Context

Proceedings of the ALTE Barcelona Conference, July 2001

Edited by
Michael Milanovic
and Cyril J Weir

Series Editors
Michael Milanovic
and Cyril J Weir

UNIVERSITY *of* CAMBRIDGE
ESOL Examinations

CAMBRIDGE
UNIVERSITY PRESS

European language testing in a global context

Proceedings of the ALTE Barcelona Conference
July 2001

Also in this series:

An investigation into the comparability of two tests of English as a Foreign Language: The Cambridge-TOEFL comparability study
Lyle F. Bachman, F. Davidson, K. Ryan, I.-C. Choi

Test taker characteristics and performance: A structural modeling approach
Antony John Kunnan

Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem
Michael Milanovic, Nick Saville

The development of IELTS: A study of the effect of background knowledge on reading comprehension
Caroline Margaret Clapham

Verbal protocol analysis in language testing research: A handbook
Alison Green

A multilingual glossary of language testing terms
Prepared by ALTE members

Dictionary of language testing
Alan Davies, Annie Brown, Cathie Elder, Kathryn Hill, Tom Lumley, Tim McNamara

Learner strategy use and performance on language tests: A structural equiation modelling approach
James Enos Purpura

Fairness and validation in language assessment: Selected papers from the 19th Language Testing research Colloquium, Orlando, Florida
Antony John Kunnan

Issues in computer-adaptive testing of reading proficiency
Micheline Chalhoub-Deville

Experimenting with uncertainty: Essays in honour of Alan Davies
A.Brown, C. Elder, N. Iwashita, E. Grove, K. Hill, T. Lumley, K. O'Loughlin, T. McNamara

An empirical investigation of the componentiality of L2 reading in English for academic purposes
Cyril Weir

The equivalence of direct and semi-direct speaking tests
Kieran O'Loughlin

A qualitative approach to the validation of oral language tests
Anne Lazaraton

Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913 – 2002
Edited by Cyril Weir and Michael Milanovic

Unpublished

The development of CELS: A modular approach to testing English Language Skills
Roger Hawkey

Testing the Spoken English of Young Norwegians: a study of testing validity and the role of 'smallwords' in contributing to pupils' fluency
Angela Hasselgren

Changing language teaching through language testing: A washback study
Liying Cheng

European language testing in a global context

Proceedings of the ALTE Barcelona Conference
July 2001

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP, UK

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2004

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times 10/12pt. *System* QuarkXPress®

A catalogue record for this book is available from the British Library

ISBN 0 521 82897 X hardback
ISBN 0 521 53587 5 paperback

Contents

Section Three
A European View

Section Four
Work in Progress

Series Editor's note

The conference papers presented in this volume represent a small subset of the many excellent presentations made at the ALTE conference – European Language Testing in a Global Context – held in July 2001 in Barcelona in celebration of the European Year of Languages – 2001. They have been selected to provide a flavour of the issues that the conference addressed. A full listing of all presentations is attached at the end of this note.

The volume is divided into three parts. The first, with two papers, one written by Charles Alderson and the other by Antony Kunnan, has a focus on more general issues in Language Testing.

Alderson looks at some key issues in the field; he considers "the shape of things to come" and asks if it will be the "normal distribution". Using this pun to structure his paper, he focuses on two aspects of language testing; the first relates to the technical aspects of the subject (issues of validity, reliability, impact etc.), the second relates to ethical and political concerns.

Most of his paper chimes well with current thinking on the technical aspects and, as he admits, much of what he presents is not new and is uncontroversial. Within the European context he refers to the influential work of the Council of Europe, especially the Common European Framework and the European Language Portfolio; he describes a number of other European projects, such as DIALANG and the national examination reform project in Hungary, and he praises various aspects of the work of ALTE (e.g. for its Code of Practice, for organising useful conferences, for encouraging exchange of expertise among its members, and for raising the profile of language testing in Europe).

In focusing on the political dimension, however, he positions himself as devil's advocate and sets out to be provocative – perhaps deliberately introducing a "negative skew" into his discussion. As always his contribution is stimulating and his conclusions are certainly controversial, particularly his criticism of ALTE and several other organisations. These conclusions would not go unchallenged by many ALTE members, not least because he misrepresents the nature of the association and how it operates.

Kunnan's paper discusses the qualities of test fairness and reflects his longstanding concerns with the issues involved in this area. The framework he presents is of great value to the field of Language Testing and Kunnan has contributed significantly to the on-going debate on the qualities of test fairness within ALTE.

The second part of the volume presents a number of research studies. Anne

Lazaraton focuses on the use of qualitative research methods in the development and validation of language tests. Lazaraton is a pioneer of qualitative research in language testing and her involvement dates back to the late eighties and early nineties when such approaches were not yet widely used in the field. It is in part due to her efforts that researchers are now more willing to embrace approaches that can provide access to the rich and deep data of qualitative research. Readers are encouraged to look at her volume in this series (*A Qualitative Approach to the Validation of Oral Language Tests*).

Vivien Berry and Jo Lewkowicz focus on the important issue of compulsory language assessment for graduating students in Hong Kong. Their paper considers alternatives to using a language test alone for this purpose and looks at the applicability of variations on the portfolio concept. Jim Purpura's work on the validation of questionnaires, which addresses the interaction of personal factors and second language test performance, represents an interesting and challenging dimension of validation in language testing. Readers may also wish to refer to Purpura's volume in this series (*Learner strategy use and performance on language tests: A structural equation modelling approach*), which looks in more depth at the development of questionnaires to determine personal factors and a methodology that can be used to investigate their interactions with test performance.

Annie Brown's paper is particularly relevant as we move towards greater use of computers in language testing. Such a move is of course fraught with issues, not least of which is the one of legibility that Brown addresses here. Her findings are interesting, giving us pause for thought and indicating, as she suggests, that more research is required. In the context of IELTS, such research is currently being conducted in Cambridge.

Barry O'Sullivan's paper attempts to model the factors affecting oral test performance, an area of particular significance in large-scale assessment. The paper is part of on-going research commissioned by the University of Cambridge Local Examinations Syndicate and it is hoped that a collection of research studies into the dimensions of oral assessment will be published in this series in due course.

Finally, Sari Luoma's paper looks at self-assessment in the context of DIALANG. The DIALANG project, also referred to in Alderson's paper, has been one of the key initiatives of the European Commission in relation to language testing. As such it has benefited from significant funding and generated much research potential.

The last two parts of the volume cover aspects of work in progress. On the one hand, Joe Shiels and Wolfgang Mackiewicz summarise aspects of the on-going work of the Council of Europe and the European Union in relation to language policy. On the other, a number of researchers bring us up-to-date with test development work largely, though not exclusively, in the context of ALTE. These papers provide the reader with a reasonable overview of what is going on in a number of European countries.

In the context of the conference reflected in this volume, it is appropriate to overview how ALTE has developed over the years and what is of particular concern to the members of ALTE at the moment.

ALTE has been operating for nearly a decade and a half. It was first formed when a few organisations, acknowledging the fact that there was no obvious forum for the discussion of issues in the assessment of one's own language as a foreign language in the European context, decided to meet with this aim in mind. The question of language assessment generally is an enormous one and dealt with in different ways by national and regional authorities throughout Europe and the world. Trying to bring together such a large and diverse community would have been a very significant task and far beyond the scope of ALTE's mission. ALTE's direct interests and aims are on a much smaller scale and it important to underline that it seeks to bring together those interested in the assessment of their own language as a foreign language. This is often in an international context, particularly with the more widely spoken languages but also in a national context, as is the case with lesser spoken languages in particular. While some ALTE members are located within ministries or government departments, others are within universities and cultural agencies. The members of ALTE are part of the international educational context and ALTE itself, as well as the members that form it, is a not-for-profit organisation. As a group, ALTE aims to provide a benchmark of quality in the particular domain in which it operates. Should ALTE's work be of relevance outside its own context, then so much the better, but ALTE does not set out to establish or police the standard for European language assessment in general.

The recent history of language testing in the European context is very mixed. In the case of English we are fortunate that there has been significant interest and research in this field in English speaking countries for many years. In relation to some other European languages this is not the case. ALTE recognises that the field of language testing in different languages will be at different stages of development and that developing a language testing capacity in the European context, albeit in a relatively narrow domain, is an on-going venture. Similarly, progress, in contexts where participants are free to walk away at any time, cannot be achieved through force or coercion but rather through involvement, greater understanding and personal commitment. ALTE operates as a capacity builder in the European context, albeit in a relatively narrow domain.

As with any association, ALTE has a Secretariat, based in Cambridge and elected by the membership. The Secretariat has a three-year term of office and is supported by a number of committees, made up from the membership, who oversee various aspects of ALTE's work. The group is too large for all members to be involved in everything and there are a number of sub-groups, organised by the members and focusing on particular areas of interest. The sub-groups are formed, reformed and disbanded as circumstances and

interests dictate, and at the moment there are several active ones. We will briefly describe the work of some of these here.

The whole of ALTE has been working for some time on the ALTE Framework which seeks to place the examinations of ALTE members onto a common framework, related closely through empirical study, to the Common European Framework. The process of placing examinations on the framework is underpinned by extensive work on the content analysis of examinations, guidelines for the quality production of examinations and empirically validated performance indicators in many European languages. This work has been supported by grants from the European Commission for many years and is now being taken forward by a number of sub-groups which are considering different domains of use such as language specifically for work purposes, for young learners or for study through the medium of a language.

A group has been established to look at the extent to which teacher qualifications in different languages can be harmonised and placed on some kind of framework. The group is not looking specifically at state organised qualifications but rather those common in the private sector for example, those offered by The Alliance Francaise, the Goethe Institute, the Cervantes Institute or Cambridge amongst others. It seeks to provide greater flexibility and mobility for the ever growing body of language teachers often qualified in one language and wishing to teach another while having their existing qualifications recognised as contributing to future ones in a more systematic way than is possible at present.

The Council of Europe has made and continues to make a substantial contribution to the teaching, learning and assessment of languages in the European context and in recent years has developed the concept of the European Language Portfolio as an aid and support to the language learning and teaching community. ALTE and the European Association for Quality Language Services have collaborated on the development of a portfolio for adults, which is now in the public domain. It is hoped that this will be a valuable aid to adult learners of languages in the European context.

An ALTE sub-group has been working with the Council of Europe and John Trim in the elaboration of a Breakthrough level which would complement the Waystage, Threshold and Vantage levels already developed. ALTE's work in this area has also been supported by the European Commission in the form of funding to a group of members from Finland, Ireland, Norway, Greece and Sweden who have a particular interest in language teaching and testing at the Breakthrough level.

Another ALTE sub-group has been working on the development of a multilingual system of computer-based assessment. The approach, which is based on the concept of computer adaptive testing, has proved highly successful and innovative, providing assessment in several European languages and recently won the European Academic Software award in 2000.

ALTE members have developed a multilingual glossary of language testing terms. Part of this work has been published in this series (*A multilingual glossary of language testing terms*) but is ongoing, and as new languages join ALTE, further versions of the glossary are being developed. The glossary has allowed language testers in about 20 countries to define language testing terms in their own language and thus contributes to the process of establishing language testing as a discipline in its own right. The European Commission has supported this work throughout.

In the early 1990s, ALTE developed a code of professional practice and work has continued to elaborate the concept of quality assurance in language testing through the development of quality assurance and quality management instruments for use initially by ALTE members. This work has been in progress for several years and is now in the hands of an ALTE sub-group. As noted above, developing the concept of quality assurance and its management has to be a collaborative venture between partners and is not prone to imposition in the ALTE context. ALTE members are aware that they carry significant responsibility and aim to continue to play a leading role in defining the dimensions of quality and how an effective approach to quality management can be implemented. This work is documented and has been elaborated in ALTE News as well as at a number of international conferences. Details are also available on the ALTE website: www.alte.org.

Members of ALTE are also concerned to measure the impact of their examinations and work has gone on in the context of ALTE to develop a range of instrumentation to measure impact on stakeholders in the test taking and using constituency. Roger Hawkey discusses the concept of impact in the context of the Lingua 2000 project in one of the papers in this volume – see contents page.

ALTE members meet twice a year and hold a language testing conference in each meeting location. This is an open event, details of which are available on the ALTE website. New ALTE members are elected by the membership as a whole. Members are either full – from countries in the European Union – or associate – from countries outside. For organisations which do not have the resources to be full or associate members or who operate in a related field, there is the option of observer status. Information on all of these categories of membership is available on the ALTE website.

Finally, following the success of the Barcelona conference ALTE has agreed to organise another international conference in 2005. Details are available on the website.

Mike Milanovic
Cyril Weir
March 03

Presentations at ALTE Conference Barcelona, 2001

Karine Akerman Sarkisian, Camilla Bengtsson and Monica Langerth-Zetterman
Uppsala University, Department of Linguistics, Sweden
Developing and evaluating web-based diagnostic testing in university language education

J. Charles Alderson
Lancaster University, Department of Linguistics and Modern English Language
The shape of things to come: Will it be the normal distribution?

José Bakx and Piet Van Avermaet
Katholieke Universiteit Nijmegen and Leuven, The Netherlands and Belgium
Certificate of Dutch as a Foreign Language: A more functional approach

David Barnwell
ITÉ
Using the Common European framework in an LSP setting: Rating the Italian, Spanish, French and German abilities of workers in the teleservices industry

Marsha Bensoussan and Bonnie Ben-Israel
University of Haifa, Department of Foreign Languages, Israel
Evaluating reading comprehension of academic texts: Guided multiple-choice summary completion

Aukje Bergsma
Citogroep, The Netherlands
NT2-CAT: Computer Adaptive Test for Dutch as a second language

Vivien Berry and Jo Lewkowicz
The University of Hong Kong, The English Centre, Hong Kong
European solutions to non-European problems

Geoff Brindley
Macquarie University, NCELTR, Australia
Investigating outcomes-based assessment

Annie Brown
The University of Melbourne, Language Testing Research Centre,
The impact of handwriting on the scoring of essays

Annie Brown (Co-authors: Noriko Iwashita and Tim McNamara)
The University of Melbourne, Language Testing Research Centre,
Investigating rater's orientations in specific-purpose task-based oral

Peter Brown and Marianne Hirtzel
EAQUALS, ALTE
The EAQUALS/ALTE European Language Portfolio

Jill Burstein and Claudia Leacock
ETS Technologies, USA
Applications in Automated Essay Scoring and Feedback

Ross Charnock
Université Paris 9 – Dauphine
Taking tests and playing scales – remarks on integrated tests

David Coniam
The Chinese University of Hong Kong, Faculty of Education, Hong Kong
Establishing minimum-language-standard benchmark tests for English language teachers in Hong Kong

Margaretha Corell and Thomas Wrigstad
Stockholm University Department of Scandinavian Languages and Centre for Research on Bilingualism.
What's the difference? Analysis of two paired conversations in the Oral examination of the National Tests of Swedish as a Second/Foreign Language

Benó Csapó and Marianne Nikolov
University of Szeged and University of Pécs, Hungary
Hungarian students' performances on English and German tests

John H.A.L. de Jong
Language Testing Services, The Netherlands
Procedures for Relating Test Scores to Council of Europe Framework

Clara Maria de Vega Santos
Universidad de Salamanca, Spain
Otra manera de aproximarse a la Evaluación: La Asociación Europea de Examinadores de Lenguas (presentation in Spanish)

Veerle Depauw and Sara Gysen
Catholic University of Leuven, Centre for Language and Migration, Belgium
Measuring Dutch language proficiency. A computer-based test for low-skilled adults and an evaluation system for primary school pupils

Urszula Dobesz
Uniwersytet Wroclawski, Poland
To know, to understand, to love (Knowledge about Poland in the teaching of Polish as a second language)

Ana Maria Ducasse
LaTrobe University, Spanish Department, Australia
Assessing paired oral interaction in an oral proficiency test

Ina Ferbězar and Marko Stabej
University of Ljubljana, Centre for Slovene as a Second/Foreign Language and Department of Slavic Languages and Literature, Slovenia
Developing and Implementing Language Tests in Slovenia

Jésus Fernández and Clara Maria de Vega Santos
Universidad de Salamanca, Spain
Advantages and disadvantages of the Vantage Level: the Spanish version (presentation in Spanish)

Neus Figueras
Generalitat de Catalunya, Department d'Ensenyament, Spain
Bringing together teaching and testing for certification. The experience at the Escoles Oficials d'Idiomes

Nicholas Gardner
Basque Government, Basque country
Test of Basque (EGA)

Malgorzata Gaszynska
Jagiellonian University, Poland
Testing Comprehension in Polish as a Foreign Language (paper in Spanish)

April Ginther (Co-author: Krishna Prasad)
Purdue University, English Department, Biloa University, and Purdue University, USA
Characteristics of European TOEFL Examinees

Giuliana Grego Bolli
Università per Stranieri di Perugia, Italy
Measuring and evaluating the competence of the Italian as a second language

Rüediger Grotjahn
Ruhr-Universitaet Bochum, Germany
TestDAF: Theoretical basis and empirical research

Anne Gutch
UCLES, UK
A major international exam: The revised CPE

H.I. Hacquebord and S.J. Andringa
University of Groningen, Applied Linguistics, The Netherlands
Testing text comprehension electronically

Roger Hawkey
c/o UCLES, UK
Progetto Lingue 2000: Impact for Language Friendly Schools

Nathalie Hirschprung
Alliance Française, France
Teacher certifications produced by the Alliance Française in the ALTE context

Maria Iakovou
University of Athens, School of Philosophy, Greece
The teaching of Greek as a foreign language: Reality and perspectives

Miroslaw Jelonkiewicz
Warsaw University and University of Wroclaw, Poland
Describing and Testing Competence in Polish Culture

Miroslaw Jelonkiewicz
Warsaw University, Poland
Describing and gauging competence in Polish culture

Neil Jones
UCLES, UK
Using ALTE Can-Do statements to equate computer-based tests across languages

Neil Jones, Henk Kuijper and Angela Verschoor
UCLES, Citogroep, UK, The Netherlands
Relationships between paper and pencil tests and computer based testing

Sue Kanburoglu Hackett and Jim Ferguson
The Advisory Council for English Language Schools Ltd, Ireland
Interaction in context: a framework for assessing learner competence in action

Lucy Katona
Idegennyelvi Továbbképzö Központ (ITK), Hungary
The development of a communicative oral rating scale in Hungary

Antony John Kunnan
Assoc. Prof., TESOL Program, USA
Articulating a fairness model

Rita Kursite
Jaunjelgava Secondary School, Latvia
Analyses of Listening Tasks from Different Points of View

Michel Laurier and Denise Lussier
University of Montreal – Faculty of Education and McGill University
The development of French language tests based on national benchmarks

Anne Lazaraton
University of Minnesota, USA
Setting standards for qualitative research in language testing

Jo Lewkowicz
The University of Hong Kong, The English Centre, Hong Kong
Stakeholder perceptions of the text in reading comprehension tests

Sari Luoma
University of Jyväskylä
Self-assessment in DIALANG

Denise Lussier
McGill University, Canada
Conceptual Framework in Teaching and Assessing Cultural Competence

Wolfgang Mackiewicz
Freie Universität Berlin, Germany
Higher education and language policy in the European Union

Waldemar Martyniuk
Jagiellonian University, Poland
Polish for Europe – Introducing Certificates in Polish as a Foreign Language

Lydia McDermott
University of Natal, Durban
Language testing, contextualised needs and lifelong learning

Debie Mirtle
MATESOL, Englishtown, Boston, USA
Online language testing: Challenges, Successes and Lessons Learned

Lelia Murtagh
ITÉ
Assessing Irish skills and attitudes among young adult secondary school leavers

Marie J. Myers
Queen's University, Canada
Entrance assessments in teacher training: a lesson of international scope

Barry O'Sullivan
The University of Reading, School of Linguistics and Applied Language Studies, UK
Modelling Factors Affecting Oral Language Test Performance: An empirical study

Silvia María Olalde Vegas and Olga Juan Lázaro
Instituto Cervantes, España
Spanish distance-learning courses: Follow-up and evaluation system

Christine Pegg
Cardiff University, Centre for Language and Communication Research, UK
Lexical resource in oral interviews: Equal assessment in English and Spanish?

Mònica Pereña & Lluís Ràfols
Generalitat de Catalunya
The new Catalan examination system and the examiners' training

Juan Miguel Prieto Hernández
Universidad de Salamanca, Spain
Problemas para elaborar y evaluar una prueba de nivel: Los Diplomas de Español como Lengua Extranjera

James E. Purpura
Columbia Teachers College, USA
Developing a computerised system for investigating non-linguistic factors in L2 learning and test performances

John Read
Victoria University of Wellington, New Zealand
Investigating the Impact of a High-stakes International Proficiency Test

Diana Rumpite
Riga Technical University, Latvia
Innovative tendencies in computer based testing in ESP

Raffaele Sanzo
Ministero della Pubblica Istruzione, Italy
Foreign languages within the frame of Italian educational reform

Joseph Sheils
Modern Languages Division, Council of Europe
Council of Europe language policy and the promotion of plurilingualism

Elana Shohamy
Tel Aviv University, School of Education, Israel
The role of language testing policies in promoting or rejecting diversity in multilingual/multicultural societies

Kari Smith
Oranim Academic College of Education, Israel
Quality assessment of Quality Learning: The digital portfolio in elementary school

M. Dolors Solé Vilanova
Generalitat de Catalunya, Centre de Recursos de Llengües Estrangeres of the Department of Education
The effectiveness of the teaching of English in the Baccalaureate school population in Catalonia. Where do we stand? Where do we want to be?

Bernard Spolsky
Bar Ilan University, Israel
Developing cross-culturally appropriate language proficiency tests for schools

Claude Springer
Université Marc Bloch, Strasbourg
A pragmatic approach to evaluating language competence: Two new certifications in France

Marko Stabej and Nataša Pirih Svetina
University of Ljubljana, Department of Slavic Languages and Literatures, Slovenia
The development of communicative language ability in Slovene as a second language

Sauli Takala, Felianka Kaftandjieva and Hanna Immonen
University of Jyväskylä, Centre for Applied Language Studies, Finland
Development and Validation of Finnish Scales of Language Proficiency

Lynda Taylor
UCLES, UK
Revising instruments for rating speaking: combining qualitative and quantitative insights

John Trim
Project Director for Modern Languages, Council of Europe
The Common European Framework of Reference for Languages and its implications for language testing

Philippe Vangeneugden and Frans van der Slik
Katholieke Universiteit Nijmegen, The Netherlands and Katholieke Universiteit Leuven, Belgium
Towards a profile related certification structure for Dutch as a foreign language. Implications of a needs analysis for profile selection and description

Juliet Wilson
UCLES, UK
Assessing Young Learners – What makes a good test?

Minjie Xing and Jing Hui Wang
Salford University, School of Languages, UK and Harbin Institute of Technology, Foreign Language Department, China
A New Method of Assessing Students' Language and Culture Learning Attitudes

Poster Presentations

Guy Bentner & Ines Quaring
Centre de Langues de Luxembourg
Tests of Luxembourgish as a foreign language

Else Lindberg & Peter Vilads Vedel
Undervisingsministeriet, Copenhagen, Denmark
The new test of Danish as a foreign language

Reidun Oanaes Andersen
Norsk Språktest, Universitetet i Bergen, Norway
Tests of Norwegian as a foreign language

José Pascoal
Universidade de Lisboa, Portugal
Tests of Portuguese as a foreign language

Heinrich Rübeling
WBT, Germany
Test Arbeitsplatz Deutsch – a workplace related language test in German as a foreign language

László Szabo
Eotvos Lorand University, Budapest, Centre for Foreign languages, Hungary
Hungarian as a foreign language examination

List of contributors

J. Charles Alderson
Department of Linguistics and Modern English Language, Lancaster University

Antony Kunnan
California State University

Anne Lazaraton
University of Minnesota, Minneapolis, USA

Vivien Berry and **Jo Lewkowicz**
University of Hong Kong

James E. Purpura
Teachers College, Columbia University

Annie Brown
University of Melbourne

Barry O'Sullivan
University of Reading

Sari Luoma
Centre for Applied Language Studies, University of Jyväskylä, Finland

Joseph Sheils
Modern Languages Division/Division des langues vivantes, DGIV Council of Europe/Conseil de l'Europe, Strasbourg

Wolfgang Mackiewicz
Conseil Européen pour les Langues/European Language Council

Rüdiger Grotjahn
Ruhr-Universität Bochum, Germany

Roger Hawkey
Educational Consultant, UK

Silvia María Olalde Vegas and **Olga Juan Lázaro**
Instituto Cervantes

Mònica Pereña and **Lluís Ràfols**
Direccío General de Política Lingüística, Departament de Cultura, Generalitat de Catalunya

Piet van Avermaet (KU Leuven), **José Bakx** (KUN), **Frans van der Slik** (KUN) and **Philippe Vangeneugden** (KU Leuven)

Nicholas Gardner
Department of Culture/Kultura Saila, Government of the Basque Country/Eusko Jaurlaritza, Spain

Giuliana Grego Bolli
Università per Stranieri di Perugia, Italy

Section 1

Issues in Language Testing

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

Plenary address to the ALTE Conference, Barcelona, July 2001

Charles Alderson
Department of Linguistics and Modern English Language
Lancaster University

Introduction

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

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

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

Washback and consequential validity

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

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

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

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

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

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

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

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

Ethics: new focus on old issues

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

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

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

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

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

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

Politics

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

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

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

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

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

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

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

Standards: codes of practice and levels

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

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

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

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

National tests

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

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

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

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

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

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

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

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

Testing and language education

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

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

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

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

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

Technology and testing

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

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

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

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

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

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

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

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

having learned something about their first performance and thus is closer to current ability.

Computers can also be user-friendly in offering a range of support to test takers: on-line Help facilities, clues, tailor-made dictionaries and more, and the use of such support can be monitored and taken into account in calculating test scores. Users can be asked how confident they are that the answer they have given is correct, and their confidence rating can be used to adjust the test score. Self-assessment and the comparison of self-assessment with test performance is an obvious extension of this principle of asking users to give insights into their ability. Similarly, adaptive tests need not be merely psychometrically driven, but the user could be given the choice of taking easier or more difficult items, especially in a context where the user is given immediate feedback on their performance. Learners can be allowed to choose which skill they wish to be tested on, or which level of difficulty they take a test at. They can be allowed to choose which language they wish to see test rubrics and examples in, and in which language results and feedback are to be presented.

An example of computer-based diagnostic tests, available over the Internet, which capitalises on the advantages I have mentioned, is DIALANG (see Chapter 8 by Sari Luoma, page 143). DIALANG uses self-assessment as an integral part of diagnosis, asking users to rate their own ability. These ratings are used in combination with objective techniques in order to decide which level of test to deliver to the user. DIALANG provides immediate feedback to users, not only on scores, but also on the relationship between their test results and their self-assessment. DIALANG also gives extensive explanatory and advisory feedback on test results. The language of administration, of self-assessment, and of feedback, is chosen by the test user from a list of 14 European languages, and users can decide which skill they wish to be tested in, in any one of 14 European languages.

One of the claimed advantages of computer-based assessment is that computers can store enormous amounts of data, including every keystroke made by candidates and their sequence and the time taken to respond to a task, as well as the correctness of the response, the use of help, clue and dictionary facilities, and much more. The challenge is to make sense of this mass of data. A research agenda is needed.

What is needed above all is research that will reveal more about the validity of the tests, that will enable us to estimate the effects of the test method and delivery medium; research that will provide insights into the processes and strategies test takers use; studies that will enable the exploration of the constructs that are being measured, or that might be measured. Alongside development work that explores how the potential of the medium might best be harnessed in test methods, support, diagnosis and feedback, we need research that investigates the nature of the most effective and meaningful

feedback; the best ways of diagnosing strengths and weaknesses in language use; the most appropriate and meaningful clues that might prompt a learner's best performance; the most appropriate use and integration of media and multimedia that will allow us to measure those constructs that might have eluded us in more traditional forms of measurement – for example, latencies in spontaneous language use, planning and execution times in task performance, speed reading and processing time more generally. And we need research into the impact of the use of the technology on learning, on learners and on the curriculum' (Alderson 2000).

In short, what *is* the added value of the use of computers?

Constructs and construct validation

Recently, language testing has come to accept Messick's conception of construct validity as being a unified, all-encompassing concept, which recognises multiple perspectives on validity. In this recently-accepted view there is No Single Answer to the validity question 'What does our test measure?' or 'Does this test measure what it is supposed to measure?'. Indeed, the question is now rephrased along the following lines: 'What is the evidence that supports particular interpretations of scores on this test?' New perspectives incorporated into this unified view of construct validity include test consequences, but there is considerable debate in the field, as I have hinted already, about whether test developers can be held responsible for test use and misuse. Is consequential validity a legitimate area of concern or a political posture?

As a result of this unified perspective, validation is now seen as on-going, as the continuous monitoring and updating of relevant information, as a process that is never complete. What is salutary and useful about this new view of construct validity is that it places the test's construct at the centre of focus, and somewhat readjusts traditional concerns with test reliability. Emphasising the centrality of constructs – what we are trying to measure – necessarily demands an applied linguistic perspective. What do we know about language knowledge and ability, and ability for use? Assessing language involves not only the technical skills and knowledge to construct and analyse a test – the psychometric and statistical side to language testing – but it also requires a knowledge of what language is, what it means to 'know' a language, and what is involved in learning a language as your mother tongue or as a second or subsequent language, what is required to get somebody to perform, using language, to the best of their ability.

In the early 1990s Caroline Clapham and I published an article in which we reported our attempt to find a consensus model of language proficiency on which we could base the revised ELTS test – the IELTS, as it became known (Alderson and Clapham 1992). We were somewhat critical of the lack of

consensus about which applied linguistic models we could use, and we reported our decision to be eclectic. If we repeated that survey now, I believe we would find more consensus: the Bachman model, as it is called, is now frequently referred to, and is increasingly influential as it is incorporated into views of the constructs of reading, listening, vocabulary and so on. The model has its origins in applied linguistic thinking by Hymes, and Canale and Swain, and by research, e.g. by Bachman and Palmer and by the Canadian Immersion studies, and it has become somewhat modified as it is scrutinised and tested. But it remains very useful as the basis for test construction, and for its account of test-method facets and task characteristics.

I have already suggested that the Council of Europe's Common European Framework will be influential in the years to come in language education generally, and one aspect of its usefulness will be its exposition of a model of language, language use and language learning – often explicitly referring to the Bachman model. The most discussed aspect by North and others of the Common European Framework to date are the various scales, which are generally perceived as the most useful aspect of the Common European Framework, not only for their characterisation and operationalisation of language proficiency and language development, but above all for their practical value in measuring and assessing learning and achievement. I would argue that the debate about what constructs to include, what model to test, has diminished in volume as testers are now engaged in exploring the empirical value and validity of these consensus models through comparability studies and the like.

Yet despite this activity, there are still many gaps in our knowledge. It is for example a matter of regret that there are no published studies of tests like the innovative Royal Society of Arts' Test of the Communicative Use of English as a Foreign Language – CUEFL (now rebadged as Cambridge's Certificates in Communicative Skills in English, UCLES 1995). Empirical studies could have thrown valuable light both on development issues – the repetition of items/tasks across different test levels, the value of facsimile texts and realistic tasks – as well as on construct matters, such as the CUEFL's relation to more traditional tests and constructs. Why the research has not been done – or reported – is probably a political matter, once again.

We could learn a lot more about testing if we had more publications from exam boards – ALTE members – about how they construct their tests and the problems they face. A recent PhD study by Sari Luoma looked at theories of test development and construct validation and tried to see how the two can come together (Luoma 2001). Her research was limited by the lack of published studies that could throw light on problems in development (the one exception was IELTS). Most published studies are sanitised corporate statements, emphasising the positive features of tests rather than discussing knotty issues in development or construct definition that have still to be

addressed. My own wish list for the future of language testing would include more accounts by developers (along the lines of Alderson, Nagy, and Öveges 2000) of how tests were developed, and of how constructs were identified, operationalised, tested and revised. Such accounts could contribute to the applied linguistic literature by helping us understand these constructs and the issues involved in operationalisation – in validating, if you like, the theory.

Pandora's boxes

Despite what I have said about the Bachman Model, McNamara has opened what he calls 'Pandora's box' (McNamara 1995). He claims that the problem with the Bachman Model is that it lacks any sense of the social dimension of language proficiency. He argues that it is based on psychological rather than socio-psychological or social theories of language use, and he concludes that we must acknowledge the intrinsically social nature of performance and examine much more carefully its interactional – i.e. social – aspects. He asks the disturbing question: 'whose performance are we assessing?' Is it that of the candidate? Or the partner in paired orals? Or the interlocutor in one-to-one tests? The designer who created the tasks? Or the rater (and the creator of the criteria used by raters)? Given that scores are what is used in reporting results, then a better understanding of how scores are arrived at is crucial. Research has intensified into the nature of the interaction in oral tests and I can confidently predict that this will continue to be a fruitful area for research, particularly with reference to performance tests.

Performance testing is not in itself a new concern, but is a development from older concerns with the testing of speaking. Only recently, however, have critiques of interviews made their mark. It has been shown through discourse analysis that the interview is only one of many possible genres of oral task, and it has become clear that the language elicited by interviews is not the same as that elicited by other types of task, and by different sorts of social interaction which do not have the asymmetrical power relations of the formal interview. Thus different constructs may be tapped by different tasks.

Hill and Parry (1992) claim that traditional tests of reading assume that texts 'have meaning', and view text, reader and the skill of reading itself as autonomous entities. In contrast, their own view of literacy is that it is socially constructed, and they see the skill of reading as being much more than decoding meaning. Rather, reading is the socially structured negotiation of meaning, where readers are seen as having social, not just individual, identities. Hill and Parry's claim is that this view of literacy requires an alternative approach to the assessment of literacy that includes its social dimension. One obvious implication of this is that what it means to understand a text will need to be revisited. In the post-modern world, where multiple

meanings are deconstructed and shown to be created and recreated in interaction, almost anything goes: what then can a text be said to 'mean' and when can a reader be said to have 'understood' a text?

Another Pandora's box

A related issue in foreign-language reading is that of levels of meaning, and the existence of reading skills that enable readers to arrive at these different levels. This hierarchy of skills has often been characterised as consisting of 'higher-order' and 'lower-order' skills, where 'understanding explicitly stated facts' is held to be 'lower-order' and 'synthesising ideas contained in text' or 'distinguishing between relevant and irrelevant ideas' is held to be 'higher-order'. However, this notion has been challenged on the grounds that, firstly, expert judges do not agree, on the whole, on whether given test questions are assessing higher- or lower-order skills, and secondly, that even for those items where experts agree on the level of skill being tested, there is no correlation between level of skill and item difficulty. Item difficulty does not necessarily relate to 'level' of skill, and students do not have to acquire lower-order skills before they can acquire higher-order skills.

Such a conclusion has proved controversial and there is some evidence that, provided that teachers can agree on a definition of skills, and provided disagreements are discussed at length, substantial agreements on matching sub-skills to individual test items (but *not* according to level of skill) can be reached. I argue that all that this research proves is that judges can be trained to agree. That does not mean that skills can be separated, or be tested separately by individual test items. Introspective accounts from students completing tests purporting to assess individual sub-skills in individual items demonstrate that students can get answers correct for the 'wrong' reason – i.e. without exhibiting the skill intended – and can get an item wrong for the right reason – i.e. whilst exhibiting the skill in question. Individuals responding to test items do so in a complex and interacting variety of different ways, and experts judging test items are not well placed to predict how students with quite different levels of language proficiency might actually respond to test items. Therefore generalisations about what skills reading test items might be testing are fatally flawed. This clearly presents a dilemma for test developers and researchers. Substantive findings on the nature of what is being tested in a reading test remain inconclusive.

A third Pandora's box

Authenticity is a long-standing concern in language testing as well as in teaching, with the oft-repeated mantra that if we wish to test and predict a candidate's ability to communicate in the real world, then texts and tasks

should be as similar to that real world as possible. More recent discussions have become more focused, but, rather like washback a few years ago, have not to date been informed by empirical research findings.

However, Lewkowicz (1997) reports a number of studies of authenticity, which result in some disturbing findings. Firstly, she found that neither native nor non-native speaker judges could identify whether listening or reading texts were or were not genuine. Secondly, students did not perceive test authenticity as an important quality that would affect their performance – they tended to be pragmatic, looking at whether they were familiar with the test format and whether they thought they would do well on the tests. Thirdly, moderating committees responsible for developing tests claimed to be authentic to target language-use situations are shown rarely to appeal to the criterion of authenticity when selecting texts and tasks on a communicative test, and frequently edit texts and tasks in order, for example, to disambiguate texts. And finally, a study of the integration of reading and writing tasks, in an attempt to make writing tasks more authentic in terms of Target Language Use needs, showed that when students were given source texts to base their writing on they did not produce writing that was rated more highly. Indeed, some students in the group that had source texts to refer to were arguably disadvantaged by copying long chunks from the source texts.

Bachman (1990) distinguishes between 'situational authenticity' and 'interactional authenticity', where situational authenticity relates to some form of replication of actual speech events in language-use situations ('life-likeness') and interactional authenticity is 'a function of the extent and type of involvement of test takers' language ability in accomplishing a test task'. Bachman argues that authenticity is not an all-or-nothing affair: a test task could be high on situational authenticity and low on interactional authenticity, or vice versa. In a later publication, Bachman and Palmer (1996) separate the notion of authenticity from that of interactiveness and define authenticity as 'the degree of correspondence of the characteristics of a given language test task to the features of a TLU task'. Bachman and Palmer consider 'authenticity' to be a 'critical quality of language tests', an aspect of what they call test usefulness, alongside validity, reliability, consequences, interactiveness and practicality, and they claim that authenticity has a strong effect on candidates' test performance. Lewkowicz's research, cited above, challenges this belief, and it is clear that much more empirical research, and less theoretical speculation, is needed before the nature and value of 'authenticity' can be resolved.

Yet another Pandora's box

Ten years ago, I argued (Alderson and North 1991) that the traditional distinction between reliability and validity, whilst conceptually distinct, is quite unclear when we examine how reliability is operationalised. It is easy

enough to understand the concept of reliability when it is presented as 'being consistent, but you can be consistently wrong'. Validity is 'being right, by measuring the ability you wish to measure', and of course this can only be measured if you do so consistently. Commentators have always argued that the two concepts are complementary, since to be valid a test needs to be reliable, although a reliable test need not be valid.

However, I argued that this is too simple. In fact, if we look at how reliability is measured, problems emerge. The classic concept of reliability is test-retest. Give a test to the same people on a second occasion and the score should remain the same. But what if they have learned from the first administration? What if their language ability has changed in some way between the two administrations? In neither case is this necessarily a matter of unreliability, but low correlations between the two administrations would be valid, *if* the ability had changed. Of course, there are obvious practical difficulties in persuading test takers to take the test again, unless there is some intervening period, but the longer the period that intervenes, the more likely we are to expect a less-than-perfect correlation between the two scores for perfectly valid reasons.

Test-retest reliability is rarely measured, not for the theoretical reasons I have presented, but for practical reasons. More common is the parallel form of reliability, where on the second occasion, a parallel form of the test is administered, in an attempt to avoid the learning effect. But immediately we are faced with the problem of the nature of parallelism. Traditional concurrent validity is the usual way of validating parallel forms of a test. So parallel form correlations would be a measure of validity, not reliability. My students are usually somewhat baffled by this 'Now you see it, Now you don't' argument that sometimes concurrent comparisons are measures of validity, other times they are reliability. And as I argue below, and argued in 1991, we *know* that it is very difficult to get high correlations between supposedly parallel tasks, let alone tests, because of the change in task characteristics or topics that has accompanied the production of the parallel version. Thus there are serious task-design reasons for believing that parallel tasks are not parallel. And complete tests, which are even more complex than single tasks, are even less likely to be parallel. In the IELTS Revision Project, we achieved correlations of about .7 or .8 between supposedly parallel reading tests, no higher than correlations between some reading tests and a test of grammar.

The most frequent form of reliability calculated is test homogeneity, which only requires one test administration and does not need parallel forms. It is usually measured by Cronbach's alpha or one of the Kuder-Richardson formulae. However, these measures explicitly assume item homogeneity: they test the hypothesis that all the items are a random sample from the same domain. But most language tests are not homogeneous and are not intended to be. We know that the text on which reading tests are based influences test

scores, and we know that different test methods will produce different measures of comprehension of the same text. This is why we advocate testing reading comprehension using multiple texts and multiple test methods. In other words, we do not expect high item correlations, and arguably a low Cronbach alpha would be a measure of the validity of our test and a high reliability coefficient would suggest that we had *not* incorporated items that were sufficiently heterogeneous.

A recent study (Swain 1993) addresses this issue from a second-language-acquisition perspective. Swain studies tests designed to measure the various aspects of communicative proficiency posited by the Canale-Swain/ Bachman family of models. In order to validate the models, the researchers wished to conduct factor analyses, which require high reliabilities of the constituent tests. However, they found remarkably low reliabilities of component test scores: scores for politeness markers correlated at .06 in requests, .18 in offers and .16 in complaints. If all component scores were added together, to get a more composite measure of reliability, the correlations between two complaints was .06, two requests .14 and two offers .18. Even when averages were computed for each student across all three speech acts and correlated with their replications – a form of split-half correlation – that coefficient was only .49. Swain comments: 'we succeeded in getting a rather low estimate of internal consistency by averaging again and again – in effect, by lengthening the test and making it more and more complex. The cost is that information on how learners' performance varies from task to task has been lost' (1993: 199).

Second-language acquisition research shows that variation in task performance will be the norm, not the exception, and it may be systematic, not random, affected by the complex interaction of various task characteristics. Both testing and task research show that performance on supposedly similar tasks varies. 'If variation in interlanguage is systematic, what does this imply about the appropriateness of a search for internal test consistency?' (op. cit. 204). Indeed one might wish to argue that a good test of second-language proficiency must have a low internal consistency.

We are thus faced with a real problem in conceptualising reliability and validity, and in knowing what statistical results to use as valid measures of test quality, be that reliability, validity or other. Indeed Swain argues that we would do well to search for 'meaningful quality criteria for the inclusion of test tasks rather than rely so heavily on a measure of internal consistency.' She cites Linn *et al.*'s suggestion (Linn, Baker and Dunbar 1991) that several such criteria might be consequences, fairness, cognitive complexity, content quality and content coverage. However, given the arguments I have put forward earlier about the complexity of washback, the difficulty of being clear about what cognitive operations are involved in responding to test tasks, the difficulty that judges have in judging test content, and the possibility that apparently unfair tests might be valid, we are clearly facing dilemmas in

operationalising both traditional criteria of reliability and validity or alternative criteria of test quality.

The real world and alternative assessment

But to come back to the real world I started in: language education, teaching and learning, and the challenge of reality. In that real world, tests are often needed for social and educational-political reasons. They are often justified in terms of the need for objectivity in selection, to avoid corruption, nepotism and favouritism. Selection on merit is often held to be a better system – for those it serves. Critics claim that tests serve the interests of the Establishment, and preserve the status quo, and recent years have seen increasing dissatisfaction with tests in some societies and calls for alternative assessment procedures and alternatives to testing.

What is usually meant by this is alternative, usually informal, procedures to traditional testing, alternatives to paper-and-pencil or computer-based tests, taken at one point in time, usually with a summative function, often high-stakes in terms of consequences, and often with claimed negative washback effects. Although alternative assessment is occasionally acknowledged to have disadvantages such as being time-consuming and difficult to administer and score, the advantages of alternative assessment are frequently proclaimed to be that they provide information that is easy for stakeholders to understand, they tend to be more integrated than traditional testing techniques and can be more easily integrated into classroom procedures. Alternative assessment procedures were developed in an attempt to make testing and assessment more responsive and accountable to individual learners, to promote learning and to enhance access and equity in education.

Of course, portfolio assessment in other subject areas is nothing new: portfolios of paintings, designs and drawings are normal in arts and graphic design, architecture and similar fields, and music students are often assessed on their ability to perform not only set pieces but also on their portfolio of self-selected pieces, performed under exam conditions but prepared and perfected over time. Nevertheless, in foreign language education, portfolios have been hailed as a major innovation, overcoming the manifold disadvantages of traditional assessment.

This is without doubt one of the latest fads in language education, and it has proved productive. On our own continent we have the European Language Portfolio, celebrated as part of the European Year of Languages, and I have no doubt that its combination of a Language Passport, a Language Biography and a Language Dossier will provide a rich record of language learners' histories, achievements and motivations. Whether it will be accepted by the educational and employment establishment is unclear. Whether alternative assessment can demonstrate its validity and reliability, its practicality and its usefulness, is also still unclear. I predict that the future will see many more attempts, not just

to persuade people with missionary zeal of the value of the European Language Portfolio and similar alternative assessments, but also critically to examine their qualities, both from the standpoint of the traditional criteria of validity and reliability, but possibly also from a different perspective. It may even be that such a critical examination of the European Language Portfolio might result in the development of different alternative criteria for assessing the quality and value of such procedures. It is conceivable that beneficial washback and motivating value might come to be better recognised and defined, with clear criteria being developed for assessing washback and motivation, rather than the present cuddly fuzziness of assertions about value.

Power and responsibility

Finally, before concluding, I should like to return to a number of my themes in order to explore their implications in the real world, and in particular the real world of ALTE and its members. I have emphasised the importance of ethical conduct in language testing, and of efforts to draw up codes of practice, in an attempt to improve standards of testing and to hold testers responsible and accountable for the quality of their tests and, to the fullest extent possible, for the impact of their activities. I have suggested that the politics of testing will receive increased attention, even though it will doubtless prove difficult to publish accounts of the influence of the agendas and actions of institutions and individuals. I believe that we need to explore these matters in a little more detail, in the context of the ALTE organisation. Jack Richards once gave a rousing speech at a TESOL convention, entitled 'The Secret Life of Methods', in which he critically examined various aspects of vested interests in TESOL, including textbook publishers, teacher-educators, examination boards, and more. That criticism was controversial but was finally published (Richards 1984). I believe that the time has come for a similar critique to be raised in the world of language testing, and these thoughts are intended to contribute to such a critique, in the interests of language testing as a profession and of the ethical conduct of language-testing practice.

Tests are, as we have seen, potentially very powerful instruments. Language is a very powerful means of communication of ideas, of beliefs, of information and more. The English language is currently a very widespread and influential language. Imagine then the power of a testing company that would have a monopoly on delivering tests of English.

The Association of Language Testers in Europe is a powerful and influential organisation. Its membership is made up of testing bodies that produce tests of their national language. Thus, CITOgroep represents the testing of Dutch, the Hungarian State Foreign Languages Examinations Board represents the testing of Hungarian, the Finnish National Certificates represent the testing of Finnish, and so on. The University of Cambridge Local Examinations Syndicate (UCLES) represents the testing of English. ALTE is

in fact a cartel, because it does not allow more than one body to represent any language at a particular level. There are occasional exceptions where tests of the full range of language ability are represented by two bodies, but that is irrelevant to the argument that ALTE is an exclusive club where only one organisation can represent any language at a given level. That in effect means that no other examining body can join ALTE to represent English, since English is covered by UCLES. And the Secretariat of ALTE is in Cambridge. Imagine the power of UCLES/ Cambridge, then, in ALTE.

But there is a dilemma. Many ALTE members also produce tests in languages other than their national language. CITOgroep, the Finnish National Language Certificates, and the Hungarian State Foreign Languages Examinations Board, to take just the three examples cited earlier, all produce tests of French, German, English and more in addition to tests of their national language. But they are officially not members of ALTE for those other languages. The question then arises of the status and indeed quality of their tests in those languages. Unscrupulous testing bodies could, if they wished, give the impression to the outside world that being a member of ALTE guarantees the quality not only of their national language exams, but also of their foreign language exams – a potentially very misleading impression indeed in the case of some ALTE members and associates.

What is the rationale for this exclusivity? Especially in an age where the power of the native speaker is increasingly questioned, where the notion of the native-speaker norm has long been abandoned in language testing, and where many organisations engage quite legitimately and properly in the testing of languages of which they are not 'native speakers'? I suggest that the notion of exclusive membership is outdated as a legitimate concept. Rather it is retained in order to ensure that UCLES be the sole provider within ALTE of exams in the most powerful language in the world at present. This is surely a commercial advantage, and I suspect, from conversations I have had, that many ALTE members are not happy with this state of affairs, and wish the playing field to be levelled. But ALTE remains a cartel.

Consider further the ALTE Code of Practice. It is an excellent document and ALTE is rightly proud of its work. But there is no enforcement mechanism; there is no way in which ALTE monitors whether its members actually adhere to the Code of Practice, and membership of ALTE is not conditional on applicants having met the standards set out in the Code. And even if they did, the quality control would presumably apply only to the tests of the national language for which the member was responsible. Thus the very existence of the ALTE Code of Practice is something of an illusion: the user and the test taker might believe that the Code of Practice is in force and guarantees the quality of the exams of ALTE members, but that is simply not true. ALTE does not operate like EAQUALS (the European Association for Quality Language Services http://www.eaquals.org), which has a rigorous inspection system that is applied to any applicant language school wishing to

join the organisation. EAQUALS can claim that its members have met quality standards. ALTE cannot.

Note that I have no doubt that ALTE has achieved a great deal. It has developed not only a Code of Practice but also a framework for test levels which, although it has now been superseded by the Council of Europe's Common European Framework, has contributed to raising international debate about levels and standards. ALTE holds very useful conferences, encourages exchange of expertise among its members, and has certainly raised the profile of language testing in Europe. But the time has come for ALTE to question its basic *modus operandi*, its conditions of membership and its role in the world of professional language testing, and to revise itself.

ALTE urgently needs also to consider its impact. Not only the impact of its own tests, which as I have already suggested ought to be researched by ALTE members. But also the impact of its very existence on societies and on non-members. In particular I am very concerned that ALTE is a powerful threat to national or local examination authorities that cannot become members. This is especially true of school-leaving examinations, which are typically developed by governmental institutions, not testing companies. Although, as I have said, many of these examinations are worthless, there are in many countries, especially in East and Central Europe, serious attempts to reform national school-leaving examinations. But there is evidence that ALTE members are operating increasingly in competition with such national bodies.

In Italy, the Progetto Lingue 2000 (http://www.cambridge-efl.org/italia/lingue2000/index.cfm) is experimenting with issuing certificates to schoolchildren of external – commercial – exams. This is bound to have an impact on the availability of locally produced, non-commercial exams. Attention ought surely to be given to enhancing the quality and currency of Italian exams of foreign languages. In Hungary, an ALTE associate member offers language certificates (of unknown quality) that are recognised by the state as equivalent to local non-commercial exams. And I understand that Poland is experimenting with a similar system: Cambridge exams will be recognised for school-leaving and university entrance purposes. But such exams are not free to the user (in Italy, they are free at present, as the educational authorities pay Cambridge direct for the entrance fees for the exams, but this is clearly unsustainable in the long term).

One ethical principle that is not mentioned by testing companies is surely that successful completion of free public education should be certified by examinations that are free to the test taker, are of high quality, and which have currency in the labour market and in higher education. This principle is undermined if expensive examinations are allowed to replace free local examination certificates.

What ALTE as a responsible professional body ought to be doing is helping to build local capacity to deliver quality examinations, regardless of whether those exam providers are members of ALTE or not. If ALTE exams replace

local exams, no local, sustainable, freely available capacity is developed, and the unaccountable influence of ALTE members spreads, to the detriment of the profession of language testing.

This is an issue of ethics, of standards, of politics and of impact – all themes I have discussed above. These are topics rarely discussed in the published literature. Nevertheless, it is time to rethink ALTE, and to recognise and develop its benefits not just for its exclusive membership, but for the profession as a whole. ALTE needs to change from being a cartel to becoming a capacity-builder.

Conclusion

To summarise first. We will continue to research washback and explore test consequences, but we will not simply describe them; rather, we will try to explain them. How does washback occur, why do teachers and students do what they do, why are tests designed and exploited the way they are and why are they abused and misused? We will burrow beneath the discourses to understand the hidden agendas of all stakeholders. I would not put any money on examination boards doing this, but I am willing to be surprised. We will develop further our codes of practice and develop ways of monitoring their implementation, not just with fine words but with action. We will develop codes of ethical principles as well and expand our understanding of their suitability in different cultural and political conditions.

The Common European Framework will grow in influence and as it is used it will be refined and enhanced; there will be an exploration of finer-grained sub-levels between the main six levels to capture better the nature of learning, and through that we will develop a better understanding of the nature of the development of language proficiency, hopefully accompanied by empirical research.

Within language education, tests will hopefully be used more and more to help us understand the nature of achievement and what goes on in classrooms – both the process and the product. The greater availability of computer-based testing, and its reduced and even insignificant cost will enhance the quality of class-based tests for placement, for progress and for diagnosis.

We will continue to research our construct. Pandora's boxes will remain open but the Model will be subject to refinement and enhancement as the nature of skills and abilities, of authenticity, and of the features of tasks are explored and are related more to what we know from second-language acquisition, just as second-language acquisition research and especially research methodology, will learn from testers.

We will continue in frustration to explore alternative assessment; the European Language Portfolio will flourish and may even be researched. We will explore an enhanced view of validity, with less stress on reliability, as we focus more and more on the individual learner, not simply on groups of test-

takers, as we try to understand performances, not just scores. But this will only happen if we do the research, if we learn from the past and build on old concerns. Developing new fads and pseudo-solutions is counter-productive, and ignoring what was written and explored ten years and more ago will not be a productive way forward. We must accumulate understanding, not pretend we have made major new discoveries.

What does all this have to do with my title?

My title is intended to focus on learning-related assessment and testing: on diagnosis, placement, progress-testing and an examination not just of why we have not made much progress in the area of achievement testing, *pace* the developments in Central Europe, but also of what the implications of test development analysis and research are for attention to low-stakes assessment that is learning- and individual-learner-related, that need not (cannot?) meet normal standards of reliability, that may not produce much variance, or that occurs where we do not expect a normal curve.

What I intend my title to emphasise is that we will be less obsessed in the future with normal distributions, with standard traditional statistics, or indeed with new statistical techniques, and more concerned to understand, by a variety of means, what it is that our tests assess, what effect they have and what the various influences on and causes of test design, test use and test misuse are. Through innovations in test research methodology, together with the opportunities afforded by computer-based testing for much friendlier test delivery, easier data handling and more fine-tuned assessment, we will get closer to the individual and closer to understanding individual performances and abilities. These will be interpreted in finer detail in relation to performance on individual tasks, which will be understood better as consisting of complexes of task characteristics, and not as an assembly of homogeneous items. The complexities of tasks, of performances and of abilities will be better appreciated and attempts will be made to understand this complexity. The Common European Framework already offers a framework within which concerted research can take place; computer-based testing can enhance test delivery and the meaning of results; and the relationship between alternative assessments like the European Language Portfolio and test-based performance can be explored and better understood.

My title also has an ambiguity: 'normal distribution' of what? Not just of test scores, but of power, of income and wealth, of access and opportunities, of expertise and of responsibilities. A better distribution of all these is also needed. We need more openness, more recognition of the quality of the work of all, more concern to build the capacity of all, not just members of an exclusive club. This, of course, will only happen if people want it to, if research is encouraged by those with the power and the money, if there is less self-interest, if there is greater co-operation and pooling of resources in a common search for understanding, and less wasteful and harmful competition and rivalry.

References

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

Alderson, J. C. 2000. Technology in testing: the present and the future. System.

Alderson, J. C., and C. Clapham. 1992. Applied linguistics and language testing: a case study. *Applied Linguistics* 13: 2, 149–67.

Alderson, J. C., C. Clapham, and D. Wall. 1995. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C., and L. Hamp-Lyons. 1996. TOEFL preparation courses: a study of washback. *Language Testing* 13: 3, 280–297.

Alderson, J. C., E. Nagy and E. Öveges (eds.). 2000. *English language education in Hungary, Part II: Examining Hungarian learners' achievements in English*. Budapest: The British Council.

Alderson, J. C., and B. North. (eds.). 1991. *Language Testing in the 1990s: The Communicative Legacy*. London: Modern English Publications in association with The British Council.

Alderson, J. C., and D. Wall. 1993. Does washback exist? *Applied Linguistics*, 14: 2, 115–129.

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

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

Bachman, L. F., and A. S. Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

Davies, A. 1997. Demands of being professional in language testing. *Language Testing* 14: 3, 328–39.

Hamp-Lyons, L. 1997. Washback, impact and validity: ethical concerns. *Language Testing* 14: 3, 295–303.

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

Hill, C., and K. Parry. 1992. The test at the gate: models of literacy in reading assessment. *TESOL Quarterly* 26: 3, 433–61.

Lewkowicz, J. A. 1997. *Investigating Authenticity in Language Testing*. Unpublished Ph.D., Lancaster University, Lancaster.

Linn, R. L., E. L. Baker and S. B. Dunbar. 1991. Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher*, 20 November, 15–21.

Luoma, S. 2001. *What Does Your Language Test Measure*? Unpublished Ph.D., University of Jyväskyla, Jyväskyla.

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

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

Messick, S. 1996. Validity and washback in language testing. *Language Testing* 13: 3, 241–56.

Richards, J. C. 1984. The secret life of methods. *TESOL Quarterly* 18: 1, 7–23.

Shohamy, E. 2001. *The Power of Tests*. London: Longman.

Swain, M. 1993. Second-language testing and second-language acquisition: is there a conflict with traditional psychometrics? *Language Testing* 10: 2, 191–207.

UCLES. 1995. *Certificate in Communicative Skills in English Handbook*. Cambridge: UCLES.

Wall, D. M. 1999. *The impact of high-stakes examinations on teaching: a case study using insights from testing and innovation theory*. Unpublished Ph.D., Lancaster University, Lancaster.

Watanabe, Y. 2001. *Does the university entrance examination motivate learners? A case study of learner interviews*. REM, 100–110.

2 Test fairness

Antony John Kunnan
California State University, USA

Abstract

The concept of test fairness is arguably the most critical in test evaluation but there is no coherent framework that can be used for evaluating tests and testing practice. In this paper, I present a Test Fairness framework that consists of the following test qualities: validity, absence of bias, access, administration, and social consequences. Prior to presenting this framework, I discuss early views of test fairness, test evaluation in practice and ethics for language testing. I conclude with some practical guidelines on how the framework could be implemented and a discussion of the implications of the framework for test development.

Introduction

The idea of test fairness as a concept that can be used in test evaluation has become a primary concern to language-testing professionals today, but it may be a somewhat recent preoccupation in the history of testing itself. Perhaps this is so because of the egalitarian view that tests and examinations were considered beneficial to society, as they helped ensure equal opportunity for education and employment and attacked the prior system of privilege and patronage. For this reason, tests and examinations have taken on the characteristic of infallibility. But everyone who has taken a test knows that tests are not perfect; tests and testing practices need to be evaluated too.

The first explicit, documented mention of a test quality was in the 19th century, after competitive examinations had become entrenched in the UK. According to Spolsky (1995), in 1858 a committee for the Oxford examinations 'worked with the examiners to ensure the general *consistency* of the examination as a whole' (p. 20). According to Stigler (1986), Edgeworth articulated the notion of *consistency* (*or reliability*) in his papers on error and chance much later, influenced by Galton's anthropometric laboratory for studying physical characteristics. As testing became more popular in the later decades of the 19th century and early 20th century, modern measurement

theory (with influential treatments from Galton, Edgeworth, Spearman, Pearson, Stevens, Guilford and Thurstone) developed techniques including correlation and factor analysis. These statistical procedures became the primary evaluative procedures for test development and test evaluation. In modern language assessment, test evaluation has clearly derived from this tradition of statistical procedures (and quantitative methods). In recent years there has been interest in using qualitative methods, and the concept of fairness too has specifically emerged, but a framework that includes these methods and concepts has not been debated.

In this paper, influenced by the work of Messick on test validation, and by ethics and philosophy, I present a framework of test fairness that broadens the scope of traditional test evaluation.

Early approaches to test evaluation

Many testing professionals hold the view that testing research has always focused on issues of fairness (and related matters such as bias, justice and equality), within the framework of test evaluation through the concepts of validity and reliability. A closer examination of this view should clarify whether this is an acceptable idea. Influenced by the work in statistics and measurement and the *Standards* (actually recommendations) for educational and psychological tests and manuals of the American Psychological Association (1954), Lado (1961) was the first author in modern language assessment to write about test evaluation in terms of validity (which encompasses face validity, validity by content, validation of the conditions required to answer the test items, and empirical validation in terms of concurrent and criterion-based validation) and reliability. Later, Davies (1968) presented a scheme for determining validities which listed five types of validity – face, content, construct, predictive and concurrent – and Harris (1969) urged test writers to establish the characteristics of a good test by examining tests in terms of content, empirical (predictive and concurrent), and face validity.

The *Standards* were reworked during this time (APA 1966, 1974) and the interrelatedness of the three different aspects of validity (content, criterion-related and construct validities) was recognised in the 1974 version. This trinitarian doctrine of content, criterion-related and construct validity (reduced in number because the concurrent and predictive validity of the 1954 version were combined and referred to as criterion-related) continued to dominate the field. In 1985, the *Standards* were reworked again and titled 'Standards for educational and psychological test' (instead of *Standards for tests*).[1] This new reworking included Messick's unified and expanded conceptual framework of validity, which was fully articulated in Messick (1989) with attention to values and social consequences of tests and testing as facets of validity of test-score interpretation. But most books in language testing did not present Messick's

unified and expanded view of validity (see Henning 1987; Hughes 1989; Alderson, Clapham and Wall 1995; Genesee and Upshur 1996; and Brown 1996). Only Bachman (1990) presented and discussed Messick's unified and expanded view of validity. Thus, validity and reliability continue to remain the dominant concepts in test evaluation, and fairness has remained outside the mainstream.

The 'Test Usefulness' approach to test evaluation

The 1990s brought a new approach to test evaluation. Translating Messick's conceptual framework, Bachman and Palmer (1996, B and P hereafter) articulated their ideas regarding test evaluation qualities: 'the most important consideration in designing and developing a language test is in the use for which it is intended, so that the most important quality of a test is its usefulness' (p. 17). They expressed their notion thus: 'Usefulness = Reliability + Construct Validity + Authenticity + Interactiveness + Impact + Practicality' (p. 18). This representation of test usefulness, they asserted, 'can be described as a function of several different qualities, all of which contribute in unique but interrelated ways to the overall usefulness of a given test' (p.18). The B and P approach does not directly address the concept of fairness but they do show an awareness of it in their use of Messick's expanded view of validity and their inclusion of impact as one of the 'test usefulness' qualities.

Test evaluation in practice

Another way of noting which test evaluation qualities were important to researchers is to examine the research they carried out. For example, researchers at Educational Testing Service, Princeton, examined the TOEFL, the TSE and the TWE, and produced 78 research and technical reports (55 on the TOEFL, 11 on TSE, and 12 on TWE). The areas of inquiry include test validation, test information, examinee performance, test use, test construction, test implementation, test reliability, and applied technology (ETS 1997).[2] The University of Cambridge Local Examinations Syndicate, UK (UCLES), which administers many EFL tests including the FCE, CPE and the IELTS, examined their tests too but judging from the IELTS research reports (IELTS 1999), the range of studies is limited to investigations of test reliability, validity and authenticity.[3 4] The English Language Institute, University of Michigan, Ann Arbor, which administers many EFL tests, produced a technical manual in support of the Michigan English Language Assessment Battery.[5] This manual includes discussions on validity and reliability using quantitative methods.[6]

The 13th edition of the *Mental Measurement Yearbook* (Impara and Blake 1998; MMY for short) has 693 reviews of 369 tests and includes reviews of 21 tests in English, 13 in reading, and four in foreign languages. Of these 38

test reviews, most of them uniformly discuss the five kinds of validity and reliability (typically, in terms of test-retest and internal consistency), and a few reviews discuss differential item functioning and bias.[7] The 10th Volume of *Test Critiques* (Keyser and Sweetland 1994) has 106 reviews which include seven related to language. Although the reviews are longer and not as constrained as the ones in the MMY, most reviews only discuss the five kinds of validity and the two kinds of reliability. The *Reviews of English Language Proficiency Tests* (Alderson *et al*. 1987) is the only compilation of reviews of English language proficiency tests available. There are 47 reviews in all and they follow the MMY's set pattern of only discussing reliability and validity, mostly using the trinitarian approach to validity while a few reviews also include discussions of practicality. There is no direct reference to test fairness.

Early test bias studies

While these general studies and reviews do not as a rule focus on the concept of fairness, a separate interest in developing culture and bias-free tests developed in educational testing. These studies began with the narrow focus on test and item-bias studies and then developed into the technical literature now known as DIF (Differential Item Functioning) studies.[8] Early landmark studies, cited in Willingham and Cole (1997), examined predictions of grades for black and white college students (Cleary 1968, and Cleary and Hilton 1969), differential validity of employment tests by race (Hunter, Schmidt and Hunter 1979), and fair selection models (Cole 1973; Cole and Moss 1989). Similarly, in language-testing research in the last two decades, gender, academic major, and native language and culture group differences have been examined the most (examples: Chen and Henning 1985; Alderson and Urquhart 1985ab; Zeidner 1986, 1987; Oltman *et al*. 1988; Hale 1988; Kunnan 1990; Ryan and Bachman 1992; Bachman *et al*. 1995; Kunnan 1995; Elder 1996; Clapham 1996, 1998; Ginther and Stevens 1998). Other relevant studies include examination of washback (Wall and Alderson 1993), test-taker feedback (Norton and Stein 1998), and test access (Brown 1993; Taylor *et al*. 1998). In summary, while some researchers are interested in test bias, the approach is fragmentary at best and not all tests are evaluated using a fairness framework.

In conclusion: first, although this overall 'engineering' approach, greatly influenced by the invention of statistical techniques, helped provide the tools necessary for validity and reliability studies, this unfortunately made most researchers complacent as they valued only statistical evidence and discounted other types of investigations and evidence. Second, single narrow-scope studies give the illusion that they have accomplished more than they set out to do. For example, a single DIF study (investigating gender differences in performances) might typically attempt to provide answers to the question of

test or item bias for certain groups, but might not be able to answer questions regarding other group differences. Or a single validation study (of, say, internal structure), while useful in its own right, would have insufficient validation evidence to claim that the test has all the desirable qualities. Third, published test reviews are narrow and constrained in such a way that none of the reviews I surveyed follow Messick's (1989) concepts of test interpretation and use and evidential and consequences bases of validation, and, therefore, they do not provide any guidance regarding these matters. In short, based on the analyses above, test evaluation is conducted narrowly and focuses mainly on validity and reliability.

Ethics in language testing

A language-test ethic has been slow to develop over the last 100 years. Spolsky (1995) convincingly argued that from the 1910s to the 1960s, social, economic and political concerns among key language-testing professions in the US (mainly at the Ford Foundation, the College Board, and Educational Test Service, Princeton) and the UK (mainly at the University of Cambridge Local Examinations Syndicate [UCLES]) dominated boardroom meetings and decisions. A language-test ethic was not evident in this period although ethical theories of different persuasions had been in existence for several centuries.

Davies (1977) was the first to make an interesting suggestion for 'test virtues' which could be seen as the first suggestion of ethical concerns in language testing. Except for Davies' 'test virtues' of reliability and validity, there has been no mention of test ethics. In the last two decades, ethical concerns emerged sporadically in language assessment. Spolsky (1981) argued that tests should be labelled, like drugs, 'Use with care'. Stevenson (1981) urged language testers to adhere to test development standards that are internationally accepted for all educational and psychological measures. Canale (1988) suggested a naturalistic-ethical approach to language testing, emphasising that language testers should be responsible for ethical use of the information they collect. Groot (1990) argued for checks on the quality of tests as essential for valid conclusions and decisions. Stansfield (1993) argued that professional standards and a code of practice are ways of bringing about ethical behaviour among testers. Alderson *et al*. (1995) reviewed principles and standards but concluded that 'language testing still lacks any agreed standards by which language tests can be evaluated, compared or selected' (p. 259).

Broadening the discussion, Cumming (1994) reviewed the functions of language assessment for recent immigrants to Canada and asked a fundamental question: 'Does the process of language assessment help or hinder…?' (p. 117). He raised three problems regarding the way in which language assessment functions within Canadian society: language assessment

may pose barriers to recent immigrants' participation in Canadian society; it may be too limited in scope; and it may put the burden of responsibility onto the performance of individual immigrants. Valdés and Figueroa (1994) addressed the reasons underlying bilingual children's poor performance on standardised tests, arguing that without an understanding of the nature of bilingualism itself, the problems encountered by bilingual individuals on such tests will continue. The last two studies began to consider language assessment from a broader, societal perspective, which features ethical issues such as justice, equity and participation.

In the last few years, momentum has gathered through publications such as the special issue of *Language Testing* guest-edited by Davies (1997a), which contained significant papers by Spolsky (1997), Lynch (1997), Hamp-Lyons (1997a), Shohamy (1997) and Davies (1997b), and other important papers by Hamp-Lyons (1997b) and Norton (1997). The International Language Testing Association (ILTA) recently published a report by the Task Force on Testing Standards (1995), which was followed by ILTA's Code of Ethics (2000) which lays out some broad guidance on how professionals should conduct themselves. However, these documents are general explorations of applied ethics and lack specific application of ethical methods which can be applied to test evaluation.

The three predominant, secular, ethical methods (utilitarianism, Kantian and deontological systems, and virtue-based ethics) may give us some guidance. To elaborate briefly, *utilitarianism*, which emphasises good results as the basis for evaluating human actions, has two main features: its teleological aspect or *consequentialist principle* and the hedonic aspect or *utility principle*. In the words of Pojman (1999), 'the consequentialist principle states that the rightness or wrongness of an act is determined by the goodness or badness of the results that flow from it. It is the end, not the means, that counts; the end justifies the means. The utility principle states that the only thing that is good in itself is some specific types of state (e.g. pleasure, happiness, welfare)' (p. 109). [9] [10]

In contrast, *Kantian* or *deontological* ethics focuses on ideals of universal law and respect for others as a basis of morality and sees the rightness or wrongness of an act in itself, not merely in the consequences of the act. In other words, the end never justifies the means. But the two kinds of deontological theory vary slightly. Act-deontologists (who are intuitionists) believe that we must consult our conscience regarding every act in order to discover whether that act is morally right or wrong. Rule-deontologists (like Kant), on the other hand, accept the notion of universal principles and believe that when we make moral decisions we are appealing to rules. [11]

Virtue-based ethics, which views moral questions from the standpoint of the moral agent with virtuous characters, has re-emerged owing to dissatisfaction that may have arisen with the previous two methods. But

virtue-based ethics calls persons to be virtuous by possessing both moral and non-moral virtues by imitation, even though there are no principles to serve as criteria for the virtues. [12]

Whatever the methodological persuasion, language-testing professionals need an ethic to support a framework of applied ethical principles that could guide professional practice. Keeping these methods in mind, we could begin to consider Hamp-Lyons' (1997b) question (with slight modification): 'What is the principle against which the ethicality of a test is to be judged?' (p. 326). Corson (1997), broadly addressing applied linguists, makes a case for the development of a framework of ethical principles by considering three principles: the principle of equal treatment, the principle of respect for persons, and the principle of benefit maximisation.

In addition to Hamp-Lyons' question cited above, we need help with other sub- or auxiliary questions, such as: What qualities should a language test have in order to be considered an ethically fair test? What are the required qualities for a language-testing practice (meaning the rights and responsibilities of all stakeholders in a test, including the test developer, test user and test taker) in order for it to be considered one with fairness or right conduct? What qualities should a code of ethics and a code of practice have so that language assessment professionals can follow ethical practice?

An ethics-inspired rationale for the Test Fairness framework

I present an ethics-inspired rationale for my test fairness framework, with a set of principles and sub-principles. The principles are based on Frankena's (1973) 'mixed deontological' system, which combines both the utilitarian and the deontological systems. Frankena suggests reconciling the two types of theory by accepting the notice of rules and principles from the deontological system but rejecting its rigidity, and by using the consequential or teleological aspect of utilitarianism but without the idea of measurement of goodness, alleviation of pain, or to bring about the greatest balance of good over evil.

Thus two general principles of justice[13] and beneficence (plus sub-principles) are articulated as follows:

Principle 1: *The Principle of Justice*: A test ought to be fair to all test takers; that is, there is a presumption of treating every person with equal respect. [14]

Sub-principle 1: A test ought to have comparable construct validity in terms of its test-score interpretation for all test takers.

Sub-principle 2: A test ought not to be biased against any test-taker groups, in particular by assessing construct-irrelevant matters.

Principle 2: *The Principle of Beneficence*: A test ought to bring about good in society; that is, it should not be harmful or detrimental to society.

Sub-principle 1: A test ought to promote good in society by providing test-score information and social impacts that are beneficial to society.

Sub-principle 2: A test ought not to inflict harm by providing test-score information or social impacts that are inaccurate or misleading.

Test fairness framework

Defining Fairness

The notion of test fairness has developed in so many ways that the various positions may appear contradictory. One useful way of understanding the many points of view is to examine recent documents that have brought this to the forefront: the Code of Fair Testing Practices in Education (1988; *Code* for short) from the Joint Committees on Testing Practices in Washington, DC and the Standards (1999, *Standards* for short) for educational and psychological testing prepared by the American Educational Research Association, American Psychological Association and the National Council on Measurement in Education.

The *Code* approach

The *Code* (1988) presents standards for educational test developers and users in four areas: developing and selecting tests, interpreting scores, striving for fairness and informing test takers. Specifically, the *Code* provides practical guidelines for test developers and users on how to strive for fairness. Keeping these guidelines in mind, standards for implementation and acceptability for the qualities are discussed here. Here is the excerpt from Section C, *Striving for Fairness,* divided into two parts, one for test developers and one for test users:

> **Test developers** should strive to make tests that are as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.
>
> **Test developers should:**
> Review and revise test questions and related materials to avoid potentially insensitive content or language.
> - Investigate the performance of test takers of different races, gender and ethnic backgrounds when samples of sufficient size are available. Enact procedures that help to ensure that differences in performance are related primarily to the skills under assessment rather than to irrelevant factors.
> - When feasible, make appropriately modified forms of tests or administration procedures available for test takers with handicapping conditions. Warn test users of potential problems in using standard norms with modified tests or administration procedures that result in non-comparable scores.

Test users should select tests that have been developed in ways that attempt to make them as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test users should:

- Evaluate the procedures used by test developers to avoid potentially insensitive content or language.
- Review the performance of test takers of different races, gender and ethnic backgrounds when samples of sufficient size are available. Evaluate the extent to which performance differences might have been caused by inappropriate characteristics of the test.
- When necessary and feasible, use appropriately modified forms of tests or administration procedures for test takers with handicapping conditions. Interpret standard norms with care in the light of the modifications that were made.

(Code 1988, p. 4–5)

The *Standards* (1999) approach

In the recent *Standards* (1999), in the chapter entitled, 'Fairness in testing and test use', the authors state by way of background that the 'concern for fairness in testing is pervasive, and the treatment accorded the topic here cannot do justice to the complex issues involved. A full consideration of fairness would explore the many functions of testing in relation to its many goals, including the broad goal of achieving equality of opportunity in our society' (p. 73). Furthermore, the document acknowledges the difficulty of defining fairness: 'the term *fairness* is used in many different ways and has no single meaning. It is possible that two individuals may endorse fairness in testing as a desirable social goal, yet reach quite different conclusions' (p. 74). With this caveat, the authors outline four principal ways in which the term is used[15]:

> The first two characterisations... relate fairness to *absence of bias* and to *equitable treatment of all examinees* in the testing process. There is broad consensus that tests should be free from bias... and that all examinees should be treated fairly in the testing process itself (e.g. afforded the same or comparable procedures in testing, test scoring, and use of scores). The third characterisation of test fairness addresses *the equality of testing outcomes* for examinee subgroups defined by race, ethnicity, gender, disability, or other characteristics. The idea that fairness requires equality in overall passing rates for different groups has been almost entirely repudiated in the professional testing literature. A more widely accepted view would hold that examinees of equal standing with respect to the construct the test is intended to measure should on average earn the same test score, irrespective of group membership... The fourth definition of fairness relates to *equity in opportunity to learn* the material covered in an

> achievement test. There would be general agreement that adequate opportunity to learn is clearly relevant to some uses and interpretations of achievement tests are clearly irrelevant to others, although disagreement might arise as to the relevance of opportunity to learn to test fairness in some specific situations.
>
> (*Standards* 1999, p. 74; emphasis added)

In addition, the document discusses two other main points: bias associated with test content and response processes, and fairness in selection and prediction. Based on these discussions, the document goes on to formulate twelve standards for fairness. The relevant standards are summarised here:

- Validity evidence collected for the whole test group should also be collected for relevant sub-groups.
- A test should be used only for the sub-groups for which evidence indicates that valid inferences can be drawn from test scores.
- When DIF exists across test-taker characteristic groups, test developers should conduct appropriate studies.
- Test developers should strive to identify and eliminate language and content that are offensive by sub-groups except when necessary for adequate representation of the domain.
- When differential prediction of a criterion for members of different sub-groups are conducted, regression equations (or appropriate equivalent) should be computed separately for each group.
- When test results are from high-stakes testing, evidence from mean score differences between relevant sub-groups should be examined and if such differences are found, an investigation should be undertaken to determine that such differences are not attributable to a source of construct under-representation or construct-irrelevance variance.

Willingham and Cole (1997) approach

Independent researchers like Willingham and Cole (1997) (in their study of gender and fair assessment) and Willingham (1999), emphasised several varying ideas in describing a system for considering fairness issues. They state that 'test fairness is an important aspect of validity... anything that reduces fairness also reduces validity... test fairness is best conceived as comparability in assessment; more specifically, comparable validity for all individuals and groups' (pp. 6–7). Using the notion of comparable validity as the central principle, Willingham suggests three criteria for evaluating the fairness of a test: 'comparability of opportunity for examinees to demonstrate relevant proficiency, comparable assessment exercises (tasks) and scores, and comparable treatment of examinees in test interpretation and use' (p. 11).

Based on these ideas, four characteristics of fairness emerge that are the most critical to fair assessment practices. They are: comparable or equitable treatment in the testing process, comparability or equality in outcomes of

learning and opportunity to learn, absence of bias in test content, language and response patterns, and comparability in selection. It is these characteristics that form the backbone of the framework that I propose below.

The Test Fairness framework

The Test Fairness framework views fairness in terms of the whole system of a testing practice, not just the test itself. Therefore, following Willingham and Cole (1997), multiple facets of fairness that includes multiple test uses (for intended and unintended purposes), multiple stakeholders in the testing process (test takers, test users, teachers and employers), and multiple steps in the test development process (test design, development, administration and use) are implicated. Thus, the model has five main qualities: validity, absence of bias, access, administration, and social consequences. Table 1 (see page 46) presents the model with the main qualities and the main focus for each of them. A brief explanation of the qualities follows:

1 **Validity**: Validity of a test score interpretation can be used as part of the test fairness framework when the following four types of evidence are collected.
   a) *Content representativeness or coverage evidence*: This type of evidence (sometimes simply described as *content validity*) refers to the adequacy with which the test items, tasks, topics and language dialect represent the test domain.
   b) *Construct or theory-based validity evidence*: This type of evidence (sometimes described as *construct validity*) refers to the adequacy with which the test items, tasks, topics and language dialect represent the construct or theory or underlying trait that is measured in a test.
   c) *Criterion-related validity evidence*: This type of evidence (sometimes described as *criterion validity*) refers to whether the test scores under consideration meet criterion variables such as school or college grades and on the job-ratings, or some other relevant variable.
   d) *Reliability*: This type of evidence refers to the reliability or consistency of test scores in terms of consistency of scores on different testing occasions (described as *stability* evidence), between two or more different forms of a test (*alternate form* evidence), between two or more raters (*inter-rater* evidence), and in the way test items measuring a construct functions (*internal consistency* evidence).

2 **Absence of bias**: Absence of bias in a test can be used as part of the test fairness framework when evidence regarding the following is collected.
   a) *Offensive content or language*: This type of bias refers to content that is offensive to test takers from different backgrounds, such as stereotypes of group members and overt or implied slurs or insults (based on gender,

race and ethnicity, religion, age, native language, national origin and sexual orientation).

b) *Unfair penalisation based on test taker's background*: This type of bias refers to content that may cause unfair penalisation because of a test taker's group membership (such as that based on gender, race and ethnicity, religion, age, native language, national origin and sexual orientation).

c) *Disparate impact and standard setting*: This type of bias refers to differing performances and resulting outcomes by test takers from different group memberships. Such group differences (as defined by salient test-taker characteristics such as gender, race and ethnicity, religion, age, native language, national origin and sexual orientation) on test tasks and sub-tests should be examined for Differential Item/Test Functioning (DIF/DTF)[16]. In addition, a differential validity analysis should be conducted in order to examine whether a test predicts success better for one group than for another. In terms of standard-setting, test scores should be examined in terms of the criterion measure and selection decisions. Test developers and users need to be confident that the appropriate measure and statistically sound and unbiased selection models are in use [17]. These analyses should indicate to test developers and test users that group differences are related to the abilities that are being assessed and not to construct-irrelevant factors.

3 Access: Access to a test can be used as part of the test fairness framework when evidence regarding the following provisions is collected.

a) *Educational access*: This refers to whether or not a test is accessible to test takers in terms of *opportunity to learn* the content and to become familiar with the types of task and cognitive demands.

b) *Financial access*: This refers to whether a test is *affordable* for test takers.

c) *Geographical access*: This refers to whether a test site is accessible in terms of distance to test takers.

d) *Personal access* here refers to whether a test provides certified test takers who have physical and/or learning disabilities with appropriate test accommodations. The 1999 *Standards* and the *Code* (1988) call for accommodation to be such that test takers with special needs are not denied access to tests that can be offered without compromising the construct being measured.

e) *Conditions or equipment access*: This refers to whether test takers are familiar with the test taking equipment (such as computers), procedures (such as reading a map), and conditions (such as using planning time).

4 Administration: Administration of a test can be used as part of the test fairness framework when evidence regarding the following conditions is collected:

a) *Physical conditions*: This refers to appropriate conditions for test administration, such as optimum light and temperature levels and facilities considered relevant for administering tests.
b) *Uniformity or consistency*: This refers to uniformity in test administration exactly as required so that there is uniformity and consistency across test sites and in equivalent forms, and that test manuals or instructions specify such requirements. Uniformity refers to length, materials and any other conditions (for example, planning time or the absence of planning time for oral and written responses) so that test takers (except those receiving accommodations due to disability) receive the test under the same conditions. Test security is also relevant to this quality, as a test's uniformity is contingent upon it being administered in secure conditions.

5 **Social consequences**: The social consequences of a test can be used as part of the test fairness framework when evidence regarding the following is collected:

a) *Washback*: This refers to the effect of a test on instructional practices, such as teaching, materials, learning, test-taking strategies, etc.
b) *Remedies*: This refers to remedies offered to test takers to reverse the detrimental consequences of a test, such as re-scoring and re-evaluation of test responses, and legal remedies for high-stakes tests. The key fairness questions here are whether the social consequences of a test and/or the testing practices are able to contribute to societal equity or not and whether there are any pernicious effects due to a particular test or testing programme [18].

In summary, these five test fairness qualities (validity, absence of bias, access, administration and social consequences), when working together, could contribute to fair tests and testing practices. Furthermore, the test fairness framework meets the guidelines of fairness in assessment contained in the recent *Code* (1988) and the *Standards* (1999). Finally, it is expected that the framework will be used in a unified manner so that a fairness argument such as the validity argument proposed by Kane (1992) can be used in defending tests as fair.

Implications for test development

The implications of the model for test development are significant. This is because the concern for fairness in language testing cannot be raised only after the test is developed and the test administered. The concern has to be present at all stages of test development: design, development, piloting and administration, and use (which includes analysis and research), although different fairness qualities may be in focus at different stages (Kunnan 2000).

Another way forward in test development is for the participants to recruit

test developers (thinkers or planners, writers, raters and researchers) from diverse groups (in terms of gender, race/ethnicity, native language, age, etc.) for training in fairness issues prior to test development. This could ensure that the many aspects of fairness would be well understood by the test developers.

Finally, the question of who is responsible for fairness-testing practices is worth raising: should it be the test developer or the test user, or both? In my view, as the two primary stakeholders of every test, both groups of individuals should be held responsible for promoting fairness.

Conclusion

In conclusion, this paper argues for a test fairness framework in language testing. This conceptualisation gives primacy to fairness and, in my view, if a test is not fair there is little value in a test having qualities such as validity and reliability of test scores. Therefore, this model consists of five interrelated test qualities: validity, absence of bias, access, administration, and social consequences.

The notion of fairness advanced here is based on the work of the *Code* (1988), the *Standards* (1999), and Willingham and Cole's (1997) notion of 'comparable validity'. This framework also brings to the forefront two qualities (*access* and *administration*) that are ignored or suppressed in earlier frameworks, as these qualities have not been seen as part of the responsibility of test developers. They have generally been delegated to test administrators and local test managers, but I propose that these two qualities should be monitored in the developmental stages and not left to the test administrators.

This framework, then, is a response to current concerns about fairness in testing and to recent discussions of applied ethics relevant to the field. Its applicability in varied contexts for different tests and testing practices in many countries would be a necessary test of its robustness. Further, I hope that the framework can influence the development of shared operating principles among language assessment professionals, so that fairness is considered vital to the professional and that societies benefit from tests and testing practices.

To sum up, as Rawls (1971) asserted, one of the principles of fairness is that institutions or practices must be *just*. Echoing Rawls, then, there is no other way to develop tests and testing practice than to make them such that primarily there is fairness and justice for all. This is especially true in an age of increasingly information-technology-based assessment, where the challenge, in Barbour's (1993) words, would be to 'imagine technology used in the service of a more just, participatory, and sustainable society on planet earth' (p. 267).

References

Alderman, D. and P. W. Holland. 1981. Item performance across native language groups on the Test of English as a Foreign Language. Princeton: Educational Testing Service.

Alderson, J. C. and A. Urquhart. 1985a. The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing* 2: 192–204.

Alderson, J. C. and A. Urquhart. 1985b. This test is unfair: I'm not an economist. In P. Hauptman, R. LeBlanc and M.B. Wesche (eds.), *Second Language Performance Testing*. Ottawa: University of Ottawa Press.

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

Alderson, J. C., C. Clapham and D. Wall. 1995. *Language Test Construction and Evaluation*. Cambridge, UK: Cambridge University Press.

American Psychological Association 1954. *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. Washington, DC. Author.

American Psychological Association 1966. *Standards for Educational and Psychological Tests and Manuals*. Washington, DC. Author.

American Psychological Association 1974. *Standards for Educational and Psychological Tests*. Washington, DC. Author.

American Psychological Association 1985. *Standards for Educational and Psychological Testing*. Washington, DC. Author.

American Psychological Association 1999. Standards for Educational and Psychological Tests. Washington, DC. Author.

Angoff, W. 1988. Validity: an evolving concept. In H. Wainer and H. Braun (eds.), *Test Validity* (pp. 19–32). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Bachman, L., F. Davidson, K. Ryan and I-C. Choi. 1995. *An Investigation into the Comparability of Two Tests of English as a Foreign Language*. Cambridge, UK: Cambridge University Press.

Bachman, L. and A. Palmer. 1996. *Language Testing in Practice*. Oxford, UK: Oxford University Press.

Barbour, I. 1993. *Ethics in an Age of Technology*. San Francisco, CA: Harper Collins.

Baron, M., P. Pettit and M. Slote (eds.). 1997. *Three Methods of Ethics*. Malden, MA: Blackwell.

Brown, A. 1993. The role of test-taker feedback in the test development process: Test takers' reactions to a rape-mediated test of proficiency in spoken Japanese. *Language Testing* 10: 3, 277–304.

Brown, J. D. 1996. *Testing in Language Programs*. Upper Saddle River, NJ: Prentice-Hall Regents.

Camilli, G. and L. Shepard. 1994. *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage.
Canale, M. 1988. The measurement of communicative competence. *Annual Review of Applied Linguistics* 8: 67–84.
Chen, Z. and G. Henning. 1985. Linguistic and cultural bias in language proficiency tests. *Language Testing* 2: 155–163.
Cizek, G. (ed.). 2001. *Setting Performance Standards*. Mahwah, NJ: Lawrence Erlbaum Associates.
Clapham, C. 1996. *The Development of IELTS*. Cambridge, UK: Cambridge University Press.
Clapham, C. 1998. The effect of language proficiency and background knowledge on EAP students' reading comprehension. In A. J. Kunnan (ed.), *Validation in Language Assessment* (pp. 141–168). Mahwah, NJ: Lawrence Erlbaum Associates.
Code of Fair Testing Practices in Education. 1988. Washington, DC: Joint Committee on Testing Practices. Author.
Corson, D. 1997. Critical realism: an emancipatory philosophy for applied linguistics? *Applied Linguistics*, 18: 2, 166–188.
Crisp, R. and M. Slote (eds.). 1997. *Virtue Ethics*. Oxford, UK: Oxford University Press.
Cumming, A. 1994. Does language assessment facilitate recent immigrants' participation in Canadian society? TESL Canada Journal 2: 2, 117–133.
Davies, A. (ed.). 1968. *Language Testing Symposium: A Psycholinguistic Approach*. Oxford, UK: Oxford University Press.
Davies, A. 1977. *The Edinburgh Course in Applied Linguistics*, *Vol. 4*. London, UK: Oxford University Press.
Davies, A. (Guest ed.). 1997a. Ethics in language testing. *Language Testing* 14: 3.
Davies, A. 1997b. Demands of being professional in language testing. *Language Testing* 14: 3, 328–339.
Elder, C. 1996. What does test bias have to do with fairness? *Language Testing* 14: 261–277.
Educational Testing Service 1997. *Program Research Review*. Princeton, NJ: Author.
Frankena, W. 1973. Ethics, 2nd ed. Saddle River, NJ: Prentice-Hall.
Genesee, F. and J. Upshur 1996. *Classroom-based Evaluation in Second Language Education*. Cambridge, UK: Cambridge University Press.
Ginther, A. and J. Stevens 1998. Language background, ethnicity, and the internal construct validity of the Advanced Placement Spanish language examination. In A. J. Kunnan (ed.), *Validation in Language Assessment* (pp. 169–194). Mahwah, NJ: Lawrence Erlbaum Associates.
Groot, P. 1990. Language testing in research and education: The need for standards. *AILA Review* 7: 9–23.

Hale, G. 1998. Student major field and text content: Interactive effects on reading comprehension in the TOEFL. *Language Testing* 5: 49–61.

Hamp-Lyons, L. 1997a. Washback, impact and validity: ethical concerns. *Language Testing* 14:3, 295–303.

Hamp-Lyons, L. 1997b. Ethics in language testing. In C. Clapham and D. Corson (eds.), *Encyclopedia of Language and Education*. (Volume 7, Language Testing and Assessment) (pp. 323–333). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Harris, D. 1969. *Testing English as a Second Language*. New York, NY: McGraw-Hill.

Henning, G. 1987. *A Guide to Language Testing*. Cambridge, MA: Newbury House.

Holland, P. and H. Wainer (eds.). 1993. *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hughes, A. 1989. *Testing for Language Teachers*. Cambridge, UK: Cambridge University Press.

Impara, J. and B. Plake (eds.). 1998. *13th Mental Measurements Yearbook*. Lincoln, NE: The Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

International English Language Testing System. 1999. *Research Reports*. Cambridge, UK: UCLES.

Kane, M. 1992. An argument-based approach to validity. *Psychological Bulletin* 112: 527–535.

Keyser, D. and R. Sweetland (eds.). 1994. *Test Critiques 10*. Austin, TX: Pro-ed.

Kim. J-O. and C. Mueller 1978. *Introduction to Factor Analysis*. Newbury Park, CA: Sage.

Kunnan, A. J. 1990. DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly* 24: 741–746.

Kunnan, A. J. 1995. *Test Taker Characteristics and Test Performance: A Structural Modelling Approach*. Cambridge, UK: Cambridge University Press.

Kunnan, A. J. 2000. Fairness and justice for all. In A. J. Kunnan (ed.), *Fairness and Validation in Language Assessment* (pp. 1–14). Cambridge, UK: Cambridge University Press.

Lado, R. 1961. *Language Testing*. London, UK: Longman.

Lynch, B. 1997. In search of the ethical test. *Language Testing* 14: 3, 315–327.

Messick, S. 1989. Validity. In R. Linn (ed.), *Educational Measurement* (pp. 13–103). London: Macmillan.

Norton, B. 1997. Accountability in language testing. In C. Clapham and D. Corson (eds.), *Encyclopedia of Language and Education*. (Volume 7, Language Testing and Assessment) (pp. 313–322). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Norton, B. and P. Stein. 1998. Why the 'monkeys passage' bombed: tests, genres, and teaching. In A. J. Kunnan (ed.), *Validation in Language Assessment*. (pp. 231–249). Mahwah, NJ: Lawrence Erlbaum Associates.

Oltman, P., J. Stricker and T. Barrows. 1988. Native language, English proficiency and the structure of the TOEFL. TOEFL Research Report 27. Princeton, NJ: Educational Testing Service.

Pojman, L. 1999. Ethics, 3rd ed. Belmont, CA: Wadsworth Publishing Co.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press.

Ross, W. 1930. *The Right and the Good.* Oxford, UK: Oxford University Press.

Ryan, K. and L. F. Bachman. 1992. Differential item functioning on two tests of EFL proficiency. *Language Testing* 9:1, 12–29.

Sen, A. and B. Williams. 1982. (eds.). *Utilitarianism and Beyond*. Cambridge, UK: Cambridge University Press.

Shohamy, E. 1997. Testing methods, testing consequences: are they ethical? Are they fair? *Language Testing* 14: 340–349.

Smart, J. and B. Williams. 1973. *Utilitarianism; For and Against*. Cambridge, UK: Cambridge University Press.

Spolsky, B. 1981. Some ethical questions about language testing. In C. Klein-Braley and D. K. Stevenson (eds.), *Practice and Problems in Language Testing 1* (pp. 5–21). Frankfurt, Germany: Verlag Peter Lang.

Spolsky, B. 1995. *Measured Words*. Oxford, UK: Oxford University Press.

Spolsky, B. 1997. The ethics of gatekeeping tests: what have we learned in a hundred years? *Language Testing* 14: 3, 242–247.

Stansfield, C. 1993. Ethics, standards, and professionalism in language testing. *Issues in Applied Linguistics* 4: 2, 189–206.

Stevenson, D. K. 1981. Language testing and academic accountability: on redefining the role of language testing in language teaching. *International Review of Applied Linguistics* 19: 15–30.

Stigler, S. 1986. *The History of Statistics*. Cambridge, MA: Belknap Press of Harvard University Press.

Taylor, C., J. Jamieson, D. Eignor and I. Kirsch. 1998. The relationship between computer familiarity and performance on computer-based TOEFL tests tasks. *TOEFL Research Report No. 61*. Princeton, NJ: Educational Testing Research.

University of Michigan English Language Institute 1996. *MELAB Technical Manual*. Ann Arbor, MI: University of Michigan Press. Author.

Valdés, G. and R. Figueroa. 1994. *Bilingualism and Testing: A Special Case of Bias*. Norwood, NJ: Lawrence Erlbaum Associates.

Wall, D. and Alderson, C. 1993. Examining washback: the Sri Lankan impact study. *Language Testing* 10: 41–70.

Willingham, W. W. and N. Cole. 1997. *Gender and Fair Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Willingham, W.W. 1999. *A systemic view of test fairness.* In S. Messick (ed.), *Assessment in Higher Education: Issues of Access, Quality, Student Development, and Public Policy* (pp. 213–242). Mahwah, NJ: Lawrence Erlbaum Associates.

Zeidner, M. 1986. Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing* 3: 80–95.

Zeidner, M. 1987. A comparison of ethnic, sex and age biases in the predictive validity of English language aptitude tests. Some Israeli data. *Language Testing* 4: 55–71.

Appendix 1

Test fairness framework

Table 1: Test fairness framework

Main quality	Main focus
1. Validity	
Content representativeness/coverage ➡	Representativeness of items, tasks, topics
Construct or theory-based validity ➡	Representation of construct/underlying trait
Criterion-related validity ➡	Test score comparison with external criteria
Reliability ➡	Stability, Alternate form, Inter-rater and Internal consistency
2. Absence of bias	
Offensive content or language ➡	Stereotypes of population groups
Unfair penalisation ➡	Content bias based on test takers' background
Disparate impact and standard setting ➡	DIF in terms of test performance; criterion setting and selection decisions
3. Access	
Educational ➡	Opportunity to learn
Financial ➡	Comparable affordability
Geographical ➡	Optimum location and distance
Personal ➡	Accommodations for test takers with disabilities
Equipment and conditions ➡	Appropriate familiarity
4. Administration	
Physical setting ➡	Optimum physical settings
Uniformity and security ➡	Uniformity and security
5. Social consequences	
Washback ➡	Desirable effects on instruction
Remedies ➡	Re-scoring, re-evaluation; legal remedies

Appendix 2

Notes

1 Angoff (1988) notes that this shift is a significant change.
2 See this document for a full listing of titles and abstracts of research studies from 1960 to 1996 for TOEFL as well as other tests such as SAT, GRE, LSAT and GMAT.
3 The FCE stands for First Certificate in English, CPE for Certificate of Proficiency in English and IELTS for International English Language Testing Service.
4 Another organisation, the Association of Language Testers of Europe (ALTE), of which UCLES is a member, has a *Code of Practice* that closely resembles the Code of Fair Testing Practices in Education (1988). However, there are no published test evaluation reports that systematically apply the *Code*.
5 MELAB stands for the Michigan English Language Assessment Battery.
6 Recent reviews in *Language Testing* of the MELAB, the TSE, the APIEL and the TOEFL CBT have begun to discuss fairness (in a limited way) along with traditional qualities such as validity and reliability.
7 This uniformity is probably also due to the way in which MMY editors prefer to conceptualise and organise reviews under headings, such as description, features, development, administration, validity, reliability and summary.
8 For DIF methodology, see Holland and Wainer (1993) and Camilli and Shepard (1994).
9 For arguments for and against utilitarianism, see Smart and Williams (1973) and Sen and Williams (1982).
10 Bentham, the classical utilitarian, invented a scheme to measure pleasure and pain called the Hedonic calculus, which registered seven aspects of a pleasurable or painful experience: intensity, duration, certainty, nearness, fruitfulness, purity and extent. According to this scheme, summing up the amounts of pleasure and pain for sets of acts and then comparing the scores could provide information as to which acts were desirable.
11 See Ross (1930) and Rawls (1971) for discussions of this system.
12 See Crisp and Slote (1997) and Baron, Pettit and Slote (1997) for discussions of virtue-based ethics. Non-secular ethics such as religion-based ethics, non-Western ethics such as African ethics, and feminist ethics are other ethical systems that may be appropriate to consider in different contexts.
13 See Rawls' (1971) *A Theory of Justice* for a clear exposition of why it is necessary to have an effective sense of justice in a well-ordered society.
14 These principles are articulated in such a way that they complement each other and if there is a situation where the two principles are in conflict, Principle 1 (The Principle of Justice) will have overriding authority. Further, the sub-principles are only explications of the principles and do not have any authority on their own.
15 The authors of the document also acknowledge that many additional interpretations of the term 'fairness' may be found in the technical testing and the popular literature.
16 There is substantial literature that is relevant to bias and DIF in language testing. For empirical studies, see Alderman and Holland (1981), Chen and Henning (1985), Zeidner (1986, 1987), Oltman *et al.* (1988), Kunnan (1990), Ryan and Bachman (1992).

17 For standard setting, the concept and practice, see numerous papers in Cizek (2001).

18 In the US, Title VII of the Civil Rights Act of 1964 provides remedies for persons who feel they are discriminated against owing to their gender, race/ethnicity, native language, national original, and so on. The Family and Education Rights and Privacy Act of 1974 provides for the right to inspect records such as tests and the right to privacy limiting official school records only to those who have legitimate educational needs. The Individuals with Disabilities Education Amendments Act of 1991 and the Rehabilitation Act of 1973 provide for the right of parental involvement and the right to fairness in testing. Finally, the Americans with Disabilities Act of 1990 provides for the right to accommodated testing. These Acts have been used broadly to challenge tests and testing practices in court.

Section 2

Research Studies

3 Qualitative research methods in language test development and validation

Anne Lazaraton
University of Minnesota, Minneapolis, MN USA

Introduction

In a comprehensive, 'state-of-the-art' article that appeared in *Language Testing*, Lyle Bachman (2000) overviews many of the ways in which language testing has matured over the last twenty years – for example, practical advances have taken place in computer-based assessment, we have a greater understanding of the many factors (e.g. characteristics of both test takers and the test-taking process) that affect performance on language tests, there is a greater emphasis on performance assessment, and there is a new concern for ethical issues in language testing. Furthermore, Bachman points to the increasing sophistication and diversity of quantitative methodological approaches in language-testing research, including criterion-referenced measurement, generalisability theory, and structural equation modelling.

In my opinion, however, the most important methodological development in language-testing research over the last decade or so has been the introduction of qualitative research methodologies to design, describe and, most importantly, to validate language tests. Bachman (2000), Banerjee and Luoma (1997), and Taylor and Saville (2001), among others, note the ways in which such qualitative research methodologies can shed light on the complex relationships that exist among test performance, test-taker characteristics and strategies, features of the testing process, and features of testing tasks, to name a few. That is, language testers have generally come to recognise the limitations of traditional statistical methods for validating language tests and have begun to consider more innovative approaches to performance test validation, approaches which promise to illuminate the assessment *process* itself, rather than just assessment outcomes.

In this paper, I explore the role of qualitative research in language test development and validation by briefly overviewing the nature of qualitative research and examining how such research can support our work as language

testers; discussing in more detail one such approach, the discourse analytic approach of conversation analysis; describing how this qualitative research method has been employed by language testers in the development and validation of ESL/EFL examinations, especially oral language tests, and by showing some recently completed work in this area and describing some practical outcomes derived from this research. Finally, I note some limitations of this type of qualitative research that should be kept in mind and point to other ways in which language testing research using qualitative methods might be conducted.

What is qualitative research?

While qualitative research has a rich tradition in the fields of education and anthropology, for example, it has a rather short history in the field of applied linguistics, which is still trying to grapple with its legitimacy. Often, in applied linguistics qualitative research is contrasted with quantitative research, as in dichotomies like the following (from Larsen-Freeman and Long 1991, p. 12):

Figure 1

Qualitative research	Quantitative research
naturalistic	controlled
observational	experimental
subjective	objective
descriptive	inferential
process-orientated	outcome-orientated
valid	reliable
holistic	particularistic
'real', 'rich', 'deep' data	hard, replicable data
ungeneralisable single case analysis	generalisable aggregate analysis

A comprehensive overview of the methodological features of interpretative qualitative research (especially ethnography) as it is conceptualised and carried out in applied linguistics can be found in Davis (1995); in a related article, Lazaraton (1995) argues that the requirements of ethnography do not adequately account for the other ten or so qualitative research traditions in existence, traditions which have different disciplinary roots, analytical goals, and theoretical motivations.

With respect to language testing, it wasn't until 1984, when Andrew Cohen proposed using a specific qualitative technique, namely introspection, to understand the testing process, that calls for a broader range of work in language testing became more frequent. At about the same time, Grotjahn

(1986) warned that a reliance on statistical analyses alone could not give us a full understanding of what a test measures, that is, its construct validity; he proposed employing more introspective techniques for understanding language tests. Fulcher (1996) observes that test designers are employing qualitative approaches more often, a positive development seeing that 'many testing instruments do not contain a rigorous applied linguistics base, whether the underpinning be theoretical or empirical. The results of validation studies are, therefore, often trivial' (p. 228). While an in-depth discussion of test validity and construct validation is beyond the scope of this paper, it should be noted that there is a growing awareness that utilising approaches from other research traditions to validate language tests is called for. For example, Kunnan (1998a) maintains that 'although validation of language (second and foreign language) assessment instruments is considered a necessary technical component of test design, development, maintenance, and research as well as a moral imperative for all stakeholders who include test developers, test-score users, test stockholders, and test takers, only recently have language assessment researchers started using a wide variety of validation approaches and analytical and interpretative techniques' (p. ix). More to the point, McNamara (1996: 85–86) argues that current approaches to test validation put too much emphasis on the individual candidate. Because performance assessment is by nature interactional, we need to pay more attention to the 'co-constructed' nature of assessment. 'In fact the study of language and interaction continues to flourish ... although it is too rarely cited by researchers in language testing, and almost not at all by those proposing general theories of performance in second language tests; this situation must change.'

Hamp-Lyons and Lynch (1998) see some reason for optimism on this point, based on their analysis of the perspectives on validity present in Language Testing Research Colloquium (LTRC) abstracts. Although they conclude that the LTRC conference is still 'positivist-psychometric dominated', it 'has been able to allow, if not yet quite welcome, both new psychometric methods and alternative assessment methods, which has led to new ways of constructing and arguing about validity' (p. 272).

Undoubtedly, there is still much to be said on this issue. Whatever approach to test validation is taken, we 'must not lose sight of what is important in any assessment situation: that decisions made on the basis of test scores are fair, because the inferences from scores are reliable and valid' (Fulcher 1999: 234). That is, 'the emphasis should always be upon the interpretability of test scores' (p. 226). In the empirical studies described in this paper, the overriding aim was to ensure confidence in just such interpretations based on the scores that the tests generated.

Specifically, it seems clear that more attention to and incorporation of discourse analysis in language test validation is needed. Fulcher (1987) remarks that 'a new approach to construct validation in which the construct

can be empirically tested can be found in discourse analysis' (p. 291). Shohamy (1991) believes that tests need to elicit more discourse and to assess such discourse carefully, and she mentions the approach of *conversation analysis* specifically as one tool for examining the interaction that takes place in oral examinations. That is, while there exists a substantial body of literature on oral language assessment, especially face-to-face interviews (much of which is reviewed in Lazaraton 2001b), until recently, research on speaking tests did not look much beyond the *outcomes* of these interviews – in most cases, the actual ratings of proficiency assigned to candidates – to the interview *process* itself, an undertaking that would allow us to 'identify and describe performance features that determine the quality of conversational interaction' in an oral interview (van Lier 1989: 497). Van Lier's seminal paper was to change all that, by stimulating an interest in undertaking empirical research into the nature of the discourse and the interaction that arises in face-to-face oral assessment. Specifically, van Lier called for studies that would even go beyond detailing the oral assessment *process*, to inform us about the turn-by-turn sequential *interaction* in the interview and whether the resulting discourse is like, or unlike, 'conversation'. In other words, a more microanalytic focus on the actual *construction* of oral test discourse by the participants would enable us to determine whether these same conversational processes are at work in the oral interview, and thus, how test interaction resembles non-test discourse.

Since that time (1989), there has been a proliferation of applied linguistics studies that analyse aspects of the discourse and interaction in oral interview situations; much work has been undertaken from a discourse-based perspective. The interested reader is urged to consult Young and He's (1998) edited volume to get an idea of just how much *has* been done in ten short years.

What is conversation analysis?

Conversation analysis (or CA, as it is often referred to) is a rigorous empirical approach to analysing talk-in-interaction with its disciplinary roots in sociology. CA employs inductive methods to search for recurring patterns across many cases without appeal to intuitive judgements of what speakers 'think' or 'mean'. The purpose of CA is to discover the systematic properties of the sequential organisation of talk, the ways in which utterances are designed to manage such sequences, and the social practices that are displayed by and embodied in talk-in-interaction. In the thirty or so years that scholars have been engaged in CA, many robust findings about the nature of native English speaker conversation have emerged; for example, the systems of turn-taking, repair, and sequence structure, are fairly well understood. There is an expanding body of CA knowledge on 'institutional talk', that is, interaction in,

for example, classrooms, courtrooms, and interviews; the role of non-verbal behaviour in conversation; the systematic properties of conversation in languages other than English; and the interactional features of talk involving non-native speakers of English. Very briefly, conversation analysis is guided by the following analytical principles:

- Using authentic, recorded data which are carefully transcribed
- Using the 'turn' as the unit of analysis
- 'Unmotivated looking' at data rather than pre-stating research questions
- Analysing single cases, deviant cases, and collections thereof
- Disregarding ethnographic and demographic particulars of the context and participants
- Eschewing the coding and quantification of data

It is my contention that conversation analysis provides a uniquely suited vehicle for understanding the interaction, and the discourse produced, in face-to-face speaking tests. With its focus on real, recorded data, segmented into turns of talk which are carefully transcribed, CA allows the analyst to move beyond mere intuitions about face-to-face test interaction and the discourse produced in it to empirically grounded findings that carry weight. Generally speaking, the conversation analyst does not formulate research questions prior to analysing data; rather, questions emerge from the data. The goal is to build a convincing and comprehensive analysis of a single case, then to search for other similar cases in order to build a collection of cases that represent some interactional phenomenon. Unlike other qualitative research approaches such as ethnography, the conversation analyst places no *a priori* importance on the sociological, demographic, or ethnographic details of the participants in the interaction or the setting in which the interaction takes place. Rather, the analyst, if interested in these issues, attempts to detect their manifestations in the discourse as it is constructed, instead of assuming some sort of 'omnirelevance' beforehand. Finally, CA studies rarely report coding or counts of data, since the emphasis in CA is on understanding single cases in and of themselves, not as part of larger aggregates of data (Schegloff (1993) presents a cogent argument for this position; for more on CA in general, see Atkinson and Heritage (1984) and Pomerantz and Fehr (1997)).

Some qualitative test validation studies

Although my own work on the Cambridge EFL Speaking Tests was primarily motivated by an inherent interest in the construction and co-construction of oral discourse, UCLES was interested in this work as a means of *construct validation*; that is, evaluating the meaningfulness and the appropriateness of interpretations that are made based on test scores. The two strands of work reported on in this paper illustrate how qualitative discourse analysis has been

used to support the test development and validation process for several of the Cambridge EFL Speaking Tests.

Very briefly, Cambridge EFL examinations are taken by more than 600,000 people in around 150 countries yearly to improve their employment prospects, to seek further education, to prepare themselves to travel or live abroad, or because they want an internationally recognised certificate showing the level they have attained in the language (UCLES 1999a). The exams include a performance testing component, in the sense that they assess candidates' ability to communicate effectively in English by producing both a written and an oral sample; these components are integral parts of the examinations. The examinations are linked to an international system of levels for assessing European languages established by the Association of Language Testers in Europe (ALTE), consisting of five user levels.

The 'Main Suite' Cambridge EFL Examinations test General English and include the *Certificate of Proficiency in English* (CPE, Level 5), the *Certificate in Advanced English* (CAE, Level 4), the *First Certificate in English* (FCE, Level 3), the *Preliminary English Test* (PET, Level 2), and the *Key English Test* (KET, Level 1). English for Academic Purposes is assessed by the *International English Language Testing System* (IELTS), jointly administered by UCLES, The British Council, and IDP Education Australia. IELTS provides proof of the language ability needed to study in English at degree level. Results are reported in nine bands, from Band 1 (Non-User) to Band 9 (Expert User). Band 6 is approximately equivalent to a good pass at Level 3 of the five-level ALTE scale. The *Cambridge Assessment of Spoken English* (CASE) was the prototype examination designed between 1990 and 1992, which has been influential in the development of the Cambridge Speaking Tests, such as FCE, IELTS, etc.

At the operational level of the speaking tests, attention to a number of test facets is evident. The Speaking Tests require that an appropriate sample of spoken English be elicited, and that the sample be rated in terms of predetermined descriptions of performance. Therefore, valid and reliable materials and criterion-rating scales, as well as a professional Oral Examiner cadre, are fundamental components in this enterprise. These concerns are addressed in the following ways:

- a paired format is employed;
- examiner roles are well defined;
- test phases are predetermined;
- a standardised format is used;
- assessment criteria are based on a theoretical model of language ability and a common scale for speaking;
- oral examiners are trained, standardised, and monitored.

A. Studies of Interlocutor Behaviour

The first set of studies I conducted focused on the behaviour of the interlocutor in the oral interview process. Cambridge Speaking Tests (with the exception of IELTS) employ both an interlocutor and an assessor as well as two candidates in a paired format. One aspect of the testing interaction that has intrigued me for many years is how interlocutors behave in this setting: how do interlocutors conduct themselves in these interviews? Are they consistent across interviews with different candidates? What sort of consistency is there between interviewers for any given speaking test? In a high-stakes EFL assessment context, these questions about consistent test delivery across candidates are of utmost importance. Thus Cambridge commissioned several studies on this topic by making available to me audiocassettes and accompanying testing materials for a number of Speaking Test administrations. These audiotapes were then carefully transcribed using CA conventions (Atkinson and Heritage 1984; see Appendix A) by graduate research assistants, who were trained in this transcription method. To engage in a conversation analysis of the transcripts, I was obliged to return to each transcript repeatedly and to scrutinise each on a turn-by-turn basis to locate and to understand the various behaviours in which the examiners engaged. While the behaviours that were identified do not exhaust the possibilities, they certainly form a group of observable actions which, once defined and identified, could be located in other transcripts, and in data from other examinations. Over the course of three studies (of CASE in 1992–93, CAE in 1994, KET in 1995, and a CAE-KET comparative study in 1996) involving 35 interlocutors and 205 candidates, it become apparent that interviewers *routinely* modified their speech (and thus test delivery) in what came to be predictable ways. For example, in (1) below, the interlocutor supplies a possible turn completion for the candidate in line 3, which the candidate then ratifies in line 4:

Supplying vocabulary

(1) CASE – Candidate 40 (2:14-18) Examiner 2

IN: is there something you don't like (.) about your job?
CA: psk .hhh ah. .hhh mm (.) yah. um nothing (.) %to- cr%
---> IN: nothing special.
CA: nothing special.

In (2), the interlocutor rephrases the initial question in lines 1 and 2 a number of times before the candidate is able to answer in line 16:

Rephrasing questions

(2) CASE – Candidate 8 (3:2-18) Examiner 2

---> IN: and do you think you will stay in the same (.) um (.)
---> area? in the same company?
(.5)
---> IN: [in the future? (.) in your job.
5 CA: [same?
CA: m-
(.)
CA: [same company?,
---> IN: [in the f- (.) in the same company in the future.
10 (1.0)
CA: uh: .hhh uh- (.) my company's eh (.) country?
---> IN: .hhh no .hhh in (.) in your future career?,=[will you stay
CA: =[hmm
---> IN: with (.) in the same area? in pharmaceuticals?
---> or do you think you may change
CA: oh: ah .hhh I want to stay (.) in the same (.) area.

The interlocutor evaluates the candidate's job by saying 'sounds interesting' in (3), a type of response not suggested by the CASE Interlocutor Frame:

Evaluating responses

(3) CASE – Candidate 37 (1:23-34) Examiner 2

IN: what's your job
CA: I'm working in an advertising agency ... our company manages psk all advertising plan? for our clients. .hhh and I'm a (.) media planner for radio there.
5 IN: media planner for [radio.
CA: [yah.
---> IN: sounds interesting?
CA: mmhmm.

In (4), the interlocutor does an embedded correction of the preposition 'in' by replacing it with the correct preposition 'to' in line 3:

Repeating and/or correcting responses

(4) CASE – Candidate 36 (2:7-10) Examiner 2

IN: which country would you like to go to.
CA: I: want to go: (.) in Spain.
---> IN: to Spain. [ah. (.) why? Spain.
CA: [yah

In (5), the interlocutor states, rather than asks, a question prescribed by the

Interlocutor Frame, 'Is English important to your career?', which requires only a confirmation from the candidate:

Stating questions that require only confirmation

(5) CASE – Candidate 9 (2:52-3:5) Examiner 2

IN: okay? .hhh an: (.) and would you like to stay with Anderson (.8) corporation? in the future?
(.)
CA: %mm I don't know%
---> IN: you don't know. [okay. .hhh um: (.) so English is
CA: [mm
---> IN: important [to your career. .hhh would you like (.) some
CA: [%yeah%
IN: higher responsibility?

Finally, in (6), the interlocutor draws a conclusion for the candidate in line 5, which the candidate then confirms and repeats as his or her own in line 6; this is an excellent example of how oral test discourse is, in fact, a co-constructed phenomenon:

Drawing conclusions for candidates

(6) CASE – Candidate 41 (2:37-42) Examiner 2

IN: %oh. I see.% .hhh (.) and will you stay in the same (.) job? with the same company? in the future? (.) do you think?
CA: hhh uh no:. .hhh hhh!
5---> IN: you want to change again.
CA: yes? [I .hhh I want to change (.) again.
IN: [hhh!

Additionally, because UCLES was primarily interested in the role of the interlocutor and the use of the Interlocutor Frame in the Speaking Tests, I devised a method for establishing how well they followed the Interlocutor Frame for each exam. For example, in the CASE study, it was determined that adherence to the frame, as worded, ranged from about 40%–100% across interlocutors. One of the benefits of the turn-by-turn analysis was being able to pinpoint specific difficulties that arose in the interaction and to analyse these instances in more detail.

What did we conclude from these studies of interlocutor behaviour? The results are clear on at least one point: the frequent and routine variations in interlocutor behaviour suggest that interlocutor cannot be considered a neutral factor in these assessments. We feel confident in saying that interlocutor language and behaviour must be standardised to some point, but it remains unclear to what extent it *should* be controlled, or to what extent it *can* be controlled. Using an interlocutor frame, monitoring interlocutor behaviour, and training examiners thoroughly are all ways of reducing, or at least

controlling, this variability. It is unlikely, however, that it would be possible, or even desirable, to eradicate the behaviour entirely, since 'the examiner factor' is the most important characteristic that distinguishes face-to-face speaking tests from their tape-mediated counterparts. Yet we should be concerned if that factor decreases test reliability, even if it appears to increase the face validity of the assessment procedure.

Finally, two practical outcomes of this work can be noted (Taylor 2001; see Taylor 2000 for further information). Firstly the findings support the utility of employing an Interlocutor Frame which guides examiners and provides candidates with consistent interlocutor input and support. At present nearly all Cambridge Speaking Tests now use an interlocutor frame. KET, PET, FCE, CAE, BEC 1/2/3 and YLE have all used a frame for some time now; an interlocutor frame will be introduced from July 2001 for IELTS and in December 2002 for the revised CPE.

Secondly, the results prompted the development of the Oral Examiner (OE) Monitoring Checklist, which has been used for 2–3 years now on a worldwide basis within the procedures for monitoring/evaluating oral examiner performance and highlighting training priorities. The form is completed by the oral examiner's Team Leader as they observe an examiner's performance with candidates in the live test situation. The form is designed to be available for use with all the Speaking Tests but is used mainly at present with KET, PET, FCE, CAE, CPE, BEC 1/2/3, and YLE. Cambridge anticipates that use of the OE Monitoring checklist will extend to the revised/new speaking tests (e.g. IELTS) over time. Results are very useful and feed back into the design of training materials and also the delivery of oral examiner training and co-ordination in each country/region.

Candidate behaviour

A second set of Cambridge-commissioned work dealt with candidate language produced on FCE and IELTS. However, because the focus of this research was solely on the features of *candidate* language, conversation analysis, which analyses dyadic interaction, could not be used. In these cases, a broader approach of discourse analysis was considered a viable option for understanding candidate speech production within the context of an oral examination.

1. Research on FCE

The study on FCE was part of a larger FCE Speaking Test Revision Project, which took place during the five years preceding the implementation of the new version in 1996. Generally, according to Taylor (1999), revisions of Cambridge EFL tests aim to account for the current target use of the candidates/learners, as well as developments in applied linguistics and description of language, in models of language-learning abilities and in

pedagogy, and in test design and measurement. The process typically begins with research, involving specially commissioned investigations, market surveys and routine test analyses, which look at performance in the five skill areas, task types, corpus use, and candidate demographics. With FCE, a survey of 25,000 students, 5,000 teachers, 1,200 oral examiners and 120 institutions asked respondents about their perspectives on the proposed revisions. This work is followed by an iterative cycle of test draft, trialling, and revision.

As a result, UCLES had a specific research question in mind for the FCE study, which was used to guide the analyses: What is the relationship between the task features in the four parts of the revised FCE Speaking Test and the candidate output in terms of speech production? The rationale for this question was to establish that the features of speech that are purported to be evaluated by the rating criteria are in fact produced by the candidates. The study consisted of two parts: first, data from the 1996 FCE Standardisation Video was analysed in order to provide supplementary information to the standardisation video materials, where appropriate, and to provide a framework for examining candidate output in a dataset of live FCE Speaking Tests that formed the basis of the second study. In the second study, Lazaraton and Frantz (1997) analysed a corpus of live data from November–December 1996 FCE test administrations. The rationale for this second project was again to establish that the features of speech which are predicted as output and which are to be evaluated by the rating criteria are actually produced by the candidates, and then to make recommendations, if necessary, about how the descriptions of the test may be amended to make the descriptions fit the likely speech output from the candidates. These two projects analysed the speech produced by 61 candidates in various administrations of FCE.

In both studies, the transcripts were analysed for the speech functions employed by the candidates in each part of the examination. This was accomplished by dividing each transcript into four parts and labelling candidate speech functions present in each section. The hypothesised speech functions that are described in the FCE materials (UCLES 1996: 26–27) were used as a starting point and were modified or supplemented as the data analysis progressed.

A total of 15 speech functions were identified in the transcripts, a number of which were ones that UCLES had identified as predicted candidate output functions. Some, however, were components of the expected functions which were thought to be too broad, too vague, or which showed too much overlap with another function. That is, then, some of the expected functions were divided into more specific functions. Finally, a few of the 15 identified functions were ones which were not predicted in the FCE materials, either directly or as part of a broader function. The analysis indicated that candidates, for the most part, did employ the speech functions that are hypothesised in the

printed FCE materials. Part 2, where candidates are required to produce a one-minute-long turn based on pictures, showed the most deviation from the expected output. In this section, UCLES hypothesised that candidates would engage in *giving information* and *expressing opinions through comparing and contrasting*. While these speech functions did occur in the data analysed, candidates also engaged in *describing, expressing an opinion, expressing a preference, justifying (an opinion, preference, choice, life decision)* and *speculating*. In fragment (7) below, the candidate spends most of her time speculating about the feelings of the people in each picture, as she is directed, but does not compare or contrast. Here is how Lazaraton and Frantz analysed this response:

(7) FCE – Candidate 43 (2:140-153) Examiner 377, Part 2
(Task: Couples: I'd like you to compare and contrast these pictures saying how you think the people are feeling)

1 yeah (.2) from the first picture I can see .hhh these two (.)
(description)
2 people they: seems not can:: cannot enjoy their .hhh their meal
(speculation)
3 (.) because these girl's face I think she's: um (think) I think
(justification)
4 she's: .hhh (.2) an- annoyed or something it's not impatient
5 and this boy: (.) she's also (.2) looks boring (.2) yeah I I
(speculation)
6 think they cannot enjoy the: this atmosphere maybe the: .hhh
(justification)
7 the:: waiter is not servings them (.) so they feel so (.) bored
(speculation)
8 or (.5) or maybe they have a argue or something like that (1.0)
9 yeah and from the second picture (.8) mmm::: this: rooms mmm:
(description)
10 looks very warm (.) and uh .hhh (.2) mmm these two people? (.)
11 they also canno- I think they are not talking to each other .hhh
(speculation)
12 they just (.) sit down over there and uh (.5) these gentleman
(description)
13 just smoking (.) yeah and this woman just look at her finger

In short, it was hoped that the results of these two studies would be useful to FCE test developers and trainers in making more accurate assessments of candidate output in the examination. Although there is no explicit task-achievement rating scheme for FCE, we suggested that the list of 15 speech functions generated from these data might prove helpful in developing one. Additionally, the list might be useful for analysing candidate output in other

Cambridge speaking tests, although it remains an empirical question whether the same functions, with the same distribution, would appear at a range of candidate ability levels. This question seems to provide fertile territory for further research.

2. Research on IELTS

Like the work on FCE, the work on IELTS should also be seen in the larger context of the IELTS Speaking Test Revision Project, recently completed, which involved clarifying task specifications to reflect input and expected candidate output, introducing an interlocutor frame to encourage standardisation of test delivery, and revising the rating scales to reflect actual candidate output (UCLES 1999b). Specifically, this study looked at the live examination performance of 20 candidates at different levels in order to identify features of language that distinguish different band scores. This required looking at linguistic features, such as grammar, vocabulary and pronunciation, as well as discourse and conversational features in order potentially to provide information for the revision of the ratings scales for IELTS Speaking. Given that there has been very little published work on the empirical relationship between candidate speech output and assigned ratings (but see Douglas 1994; Fulcher 1996, and Young 1995), there was little guidance on how to approach this task.

To illustrate the sorts of findings that emerged from this study, fragments (8)–(11) below show some interesting differences in candidate ability to use cohesive markers. For instance, candidates at the lowest band score 3 made only limited use of conjunctions. Here, the candidate uses a listing strategy in conjunction with 'and':

(8) IELTS – Candidate 9 (3:79-82) Part 2

E: okay what about the Olympic Games are you interested in that kind of sport or
C: mm yeah.
E: yeah
C: Yes I inlested in ah **football** mm **and swimming tennis** yeah.

However, 'and' above and the 'but' and 'because' below are not used to link sentences, per se:

(9) IELTS – Candidate 9 (4:92) Part 1

C: **but** I dink soccer is better

(10) IELTS – Candidate 9 (1:23-26) Part 1

E: what do you like most Sydney or Bangkok?
C: Sydney.
E: why?
C: **because** mm in Sydney it have (a rarely) traffic jam

In contrast, a candidate at a higher band score of 7 readily used cohesive ties to connect sentences to produce extended discourse:

(11) IELTS – Candidate 20 (7:166-172) Part 4

C: in one week I leave ah Australia **but** I intend to stay in Australia in Jinuree. **so** I will um I applied a for um to to participate ah in a course here at the university **and** um **if** I get in into it into this course I will come back **but at first** I have to return to Switzerland to to join the army.

In sum, this sort of structural analysis can be used to isolate and evaluate specific textual features of interest. The findings from this study proved useful by suggesting a number of modifications to the rating scales, which should not only increase the reliability of the assessment (e.g. so that all raters are clear on what 'cohesive features' are), but lead to rating scales that more accurately reflect facts about language in use. Thus confidence in score interpretations, the fundamental goal of construct validation, can undoubtedly be increased.

In fact, as Taylor and Jones (2001) report, these results were implicated in the revision of the existing holistic rating scale for IELTS into four sub-scales: pronunciation, fluency and coherence, grammatical range and accuracy, and lexical resource. Furthermore, they note that 'qualitative feedback from raters in the early stages of trialling and during the multiple rating study has proved invaluable in helping to inform production of IELTS examiner training and standardisation materials' as well as leading to further investigation of 'examiners' experience as they simultaneously deliver the test and rate candidate performance in real time' (p. 11).

A second outcome of these studies on FCE and IELTS was the realisation that, while this sort of work, using CA, is valuable, the transcription process is much too complex to generate enough data to be usable for test analysis and validation (Saville 2000). As a result, Saville and O'Sullivan (2000) developed what they refer to as 'observation checklists' for assessing candidate output on speaking tests. This was accomplished by reviewing literature in language testing and second-language acquisition to generate a list of speech functions that would be likely to be employed in the speaking tests; three types of functions were identified: informational functions, interactional functions, and functions that are used in managing interaction. Through a developmental process, these checklists have become operational and have been used, although not systematically, to validate speaking tests and evaluate task design and formats. Cambridge would also like to explore how an amended version of the observation checklist, suitable for a one-to-one rather than a paired test format, could be developed to help validate the revised format of the IELTS speaking test (Taylor 2001).

Discussion

To summarise, I believe that discourse analysis offers us a method with which we can analyse the interaction that takes place in face-to-face oral assessment, which, until very recently, was overlooked in the test validation process. These various studies on different Cambridge EFL Speaking Tests have informed us about how interlocutors and candidates behave during the test, and how this behaviour approximates to conversation. The IELTS study attempted to compare features of the discourse produced with assigned band scores, and along with the FCE studies, made some headway on informing rating scale construction and validation. The unique contribution of all these studies to the field of language testing, though, is their demonstration of both the applicability and the suitability of conversation/discourse analysis for understanding the process of oral assessment via an examination of the discourse produced in this context. As Lazaraton (2001b: 174) remarks, 'Conversation analysis has much to recommend it as a means of validating oral language tests … Perhaps the most important contribution that CA can make … is in the accessibility of its data and the claims based on them. That is, for many of us … highly sophisticated statistical analyses … are comprehensible only to those versed in those analytic procedures … The results of CA are patently observable, even if one does not agree with the conclusions at which an analyst may arrive. As such, language testers who engage in conversation analyses of test data have the potential to reach a much larger, and less exclusive readership.'

However, it is as well to keep in mind a number of shortcomings of the conversation analytic approach. Aside from a number of theoretical and conceptual objections reviewed in Lazaraton (2001b), CA is not helpful for analysing monologic data, where there is an absence of interaction; for the same reason, CA cannot be applied to the modality of writing. Equally troubling is the fact that CA is difficult, if not impossible, to learn without the benefit of tutelage under a trained analyst and/or with others. As a result, the number of researchers who feel comfortable with and driven to use the methodology will undoubtedly remain small. Related to this is the labour-intensive nature of CA, which makes it impractical for looking at large data-sets.

Another problematic aspect of qualitative research in general is that there is no clear consensus on how such research should be evaluated. Leaving aside issues of good language testing practice, discussed recently by Taylor and Saville at the 2001 AAAL conference, in the larger field of qualitative research in the social sciences and humanities there is very little agreement on the need for permanent or universal criteria for judging qualitative research. Garratt and Hodkinson (1998: 515–516) consider the two basic questions of 'criteriology', but come to no conclusions: 1) Should we be striving to

establish agreed, universal, pre-ordained criteria, against which any piece of research writing should be judged? and 2) If so, how do we choose which of the various lists of criteria advanced in the literature we should use in order to do that job? Clearly, this is a pressing issue for discussion, not just for language testing but for the larger field of applied linguistics (see Lazaraton 2001a for more on this topic).

Finally, a major obstacle is that this sort of work, which results in copious amounts of transcribed talk but no 'hard' statistical data, has generally had a difficult time finding a home in professional dissemination outlets, such as conferences and journals. Until the results of this research can be more widely circulated, argued, and replicated, the approach itself may retain its marginal status in language testing.

One final point that should be made is that other qualitative approaches to test validation also show great promise. Banerjee and Luoma (1997) overview many of these approaches, which they see as providing valuable information on test content, the properties of testing tasks, and the processes involved in taking tests and assessing test output. That is, qualitative validation techniques help to clarify the nature of performance that scores are based on, rather than to detail the psychometric properties of tests and items, as quantitative techniques are designed to do. Much of their chapter is devoted to explaining the utility of verbal reports, a topic which has received book-length coverage recently by Allison Green (1998) and is currently finding its way into empirical studies on language tests (e.g. Meiron 1999). Banerjee and Luoma also note that observations (of item writing meetings, rating sessions), questionnaires and interviews (as in Hill's 1998 study of test-taker impressions of access:), and analyses of test language (as in Lazaraton 2001b) are just a few means by which test validation can be achieved qualitatively; they believe that we can go even further in this direction, by using, for example, learner/rater diaries and qualitative software analysis programs, and by applying these techniques to look into the interpretations and uses of test scores by teachers and administrators – stakeholders whose voices often remain unheard in the validation process (Hamp-Lyons and Lynch 1998).

Conclusion

Language testing is clearly in the midst of exciting changes in perspective. It has become increasingly evident that the established psychometric methods for validating oral language tests are effective but limited, and other validation methods are required for us to have a fuller understanding of the language tests we use. I have argued that conversation analysis represents one such solution for these validation tasks. McNamara (1997: 460) sees much the same need, as he states rather eloquently: 'Research in language testing cannot consist only of a further burnishing of the already shiny chrome-plated quantitative

armour of the language tester with his (too often) sophisticated statistical tools and impressive n-size'; what is needed is the 'inclusion of another kind of research on language testing of a more fundamental kind, whose aim is to make us fully aware of the nature and significance of assessment as a social act'. As the field of language testing further matures, I am optimistic that we can welcome those whose interests and expertise lie outside the conventional psychometric tradition: qualitative researchers like myself, of course, but also those who take what Kunnan (1998b) refers to as 'postmodern' and 'radical' approaches to language assessment research. Furthermore, I would also hope, along with Hamp-Lyons and Lynch, that the stakeholders in assessment, those that use the tests that we validate, would have a greater voice in the assessment process in order to ensure that our use of test scores is, first and foremost, a responsible use.

Notes

This chapter is a slightly revised version of a plenary paper given at the ALTE conference in Barcelona, July 2001. Portions of this chapter also appear in Lazaraton (2001b).

References

Atkinson, J. M. and J. Heritage (eds.). 1984. *Structures of Social Action: Studies in Conversation Analysis*. Cambridge: Cambridge University Press.

Bachman, L. F. 2000. Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing* 17: 1–42.

Banerjee, J. and S. Luoma. 1997. Qualitative approaches to test validation. In C. Clapham and D. Corson (eds.), *Encyclopedia of Language and Education, Volume 7: Language Testing and Assessment* (pp. 275–287). Amsterdam: Kluwer.

Cohen, A. 1984. On taking language tests: What the students report. *Language Testing* 1: 70–81.

Davis, K. A. 1995. Qualitative theory and methods in applied linguistics research. *TESOL Quarterly 29*: 427–453.

Douglas, D. 1994. Quantity and quality in speaking test performance. *Language Testing 11*: 125–144.

Fulcher, G. 1987. Tests of oral performance: The need for data-based criteria. *ELT Journal 41*: 4, 287–291.

Fulcher, G. 1996. Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing 13*: 2, 208–238.

Fulcher, G. 1999. Assessment in English for Academic Purposes: Putting content validity in its place. *Applied Linguistics 20*: 2, 221–236.

Garratt, D. and P. Hodkinson. 1998. Can there be criteria for selecting research criteria? – A hermeneutical analysis of an inescapable dilemma. *Qualitative Inquiry 4*: 515–539.

Green, A. 1998. V*erbal Protocol Analysis in Language Testing Research: A Handbook*. Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate.

Grotjahn, R. 1986. Test validation and cognitive psychology: Some methodological considerations. *Language Testing 3*: 159–185.

Hamp-Lyons, L. and B. K. Lynch. 1998. Perspectives on validity: An historical analysis of language testing conference abstracts. In A. Kunnan (ed.), *Validation in Language Assessment: Selected Papers from the 17th Language Testing Research Colloquium, Long Beach* (pp. 253–276). Mahwah, NJ: Lawrence Erlbaum.

Hill, K. 1998. The effect of test-taker characteristics on reactions to and performance on an oral English proficiency test. In A. Kunnan (ed.), *Validation in Language Assessment: Selected Papers from the 17th Language Testing Research Colloquium, Long Beach* (pp. 209–229). Mahwah, NJ: Lawrence Erlbaum Associates.

Kunnan, A. (1998a). Preface. In A. Kunnan (ed.), *Validation in Language Assessment: Selected Papers from the 17th Language Testing Research Colloquium, Long Beach* (pp. ix–x). Mahwah, NJ: Lawrence Erlbaum Associates.

Kunnan, A. 1998b. Approaches to validation in language assessment. In A. Kunnan (ed.), *Validation in Language Assessment: Selected Papers from the 17th Language Testing Research Colloquium, Long Beach* (pp. 1–16). Mahwah, NJ: Lawrence Erlbaum Associates.

Larsen-Freeman, D. and M. H. Long. 1991. *An Introduction to Second Language Acquisition Research*. London: Longman.

Lazaraton, A. 1995. Qualitative research in applied linguistics: A progress report. *TESOL Quarterly 29*: 455–472.

Lazaraton, A. 2001a. *Standards in qualitative research: Whose standards? And whose research?* Paper presented at the Setting Standards for Qualitative Research in Applied Linguistics Colloquium, American Association of Applied Linguistics Annual Conference, St. Louis MO: February.

Lazaraton, A. 2001b. *A Qualitative Approach to the Validation of Oral Language Tests*. Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate.

Lazaraton, A. and R. Frantz. 1997. *An analysis of the relationship between task features and candidate output for the Revised FCE Speaking Examination.* Report prepared for the EFL Division, University of Cambridge Local Examinations Syndicate, Cambridge, UK.

McNamara, T. F. 1996. *Measuring Second Language Performance*. London: Longman.

McNamara, T. F. 1997. 'Interaction' in second language performance assessment: Whose performance?. *Applied Linguistics* 18: 446–466.

Meiron, B. E. 1999. *Inside raters' heads: An exploratory triangulated study of oral proficiency raters' thought processes*. Paper presented at the 33rd Annual TESOL Convention, New York, NY: March.

Pomerantz, A. and B. J. Fehr. 1997. Conversation analysis: An approach to the study of social action as sense making practices. In T. A. van Dijk (ed.), *Discourse as Social Action, Discourse Studies: A Multidisciplinary Introduction Volume 2* (pp. 64–91). London: Sage.

Saville, N. 2000. *Using observation checklists to validate speaking-test tasks.* Research Notes 2 (August), pp. 16–17, University of Cambridge Local Examinations Syndicate.

Saville, N. and B. O'Sullivan. 2000. *Developing observation checklists for speaking-tests*. Research Notes 3 (November): pp. 6–10, University of Cambridge Local Examinations Syndicate.

Schegloff, E. A. 1993. Reflections on quantification in the study of conversation. *Research on Language and Social Interaction* 26: 199–128.

Shohamy, E. 1991. Discourse analysis in language testing. *Annual Review of Applied Linguistics* 11: 115–131.

Taylor, L. 1999. *Constituency matters: Responsibilities and relationships in our test community*. Paper presented at the Language Testing Forum, Edinburgh: November.

Taylor, L. 2000. *Issues in speaking assessment research*. Research Notes 1 (March): pp. 8–9, University of Cambridge Local Examinations Syndicate.

Taylor, L. 2001. Personal e-mail communication, 6/7/01.

Taylor, L. and N. Jones. 2001. *Revising the IELTS speaking test.* Research Notes 4 (February): pp. 9–12, University of Cambridge Local Examinations Syndicate.

Taylor, L. B. and N. Saville. 2001. *The role of qualitative methods in setting and maintaining standards in language assessment*. Paper presented at the Setting Standards for Qualitative Research in Applied Linguistics Colloquium, American Association of Applied Linguistics Annual Conference, St. Louis, MO: February.

University of Cambridge Local Examinations Syndicate. 1996. *Notes to Accompany the 1996 FCE Standardisation Video*. Cambridge: Author.

University of Cambridge Local Examinations Syndicate. 1999a. *Guidelines to Accompany the EFL Oral Examiner Monitoring Checklist.* Cambridge: Author.

University of Cambridge Local Examinations Syndicate. 1999b. *IELTS: Annual Review 1998/9.* Cambridge: Author.

van Lier, L. 1989. Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly 23*: 489–508.

Young, R. 1995. Discontinuous language development and its implications for oral proficiency rating scales. *Applied Language Learning 6*: 13–26.

Young, R. and A. W. He (eds.). 1998. *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*. Philadelphia: John Benjamins.

Appendix 1

Transcription Notation Symbols (from Atkinson and Heritage 1984)

1. **unfilled pauses or gaps** – periods of silence, timed in tenths of a second by counting 'beats' of elapsed time. Micropauses, those of less than .2 seconds, are symbolised (.); longer pauses appear as a time within parentheses: (.5) is five tenths of a second.
2. **colon (:)** – a lengthened sound or syllable; more colons prolong the stretch.
3. **dash (-)** – a cut-off, usually a glottal stop.
4. **.hhh** – an inbreath; **.hhh!** – strong inhalation.
5. **hhh** – exhalation; **hhh!** – strong exhalation.
6. **hah, huh, heh, hnh** – all represent laughter, depending on the sounds produced. All can be followed by an (!), signifying stronger laughter.
7. **(hhh)** – breathiness within a word.
8. **punctuation:** markers of intonation rather than clausal structure; a full point **(.)** is falling intonation, a question mark **(?)** is rising intonation, a comma **(,)** is continuing intonation. A question mark followed by a comma **(?,)** represents rising intonation, but is weaker than a **(?).** An exclamation mark **(!)** is animated intonation.
9. **equal sign (=)** – a latched utterance, no interval between utterances.
10. **brackets ([ ])** – overlapping talk, where utterances start and/or end simultaneously.
11. **per cent signs (% %)** – quiet talk.
12. **asterisks (* *)** – creaky voice.
13. **carat (^)** – a marked rising shift in pitch.
14. **arrows (> <)** – the talk speeds up, arrows **(< >)** – the talk slows down.
15. **psk** – a lip smack, **tch** – a tongue click.
16. **underlining** or **CAPS** – a word or SOund is emphasised.

17. **arrow (--->)** – a feature of interest to the analyst.
18. **empty parentheses ( )** – transcription doubt, uncertainty; words within parentheses are uncertain.
19. **double parentheses (( ))** – non-vocal action, details of scene.

4 European solutions to non-European problems

Vivien Berry and Jo Lewkowicz
The University of Hong Kong

Abstract

This paper examines the major issues relating to the proposed introduction of compulsory assessment of English language proficiency for students prior to graduation from tertiary education in Hong Kong. It looks at the on-going debate relating to the introduction of a language exit-test and considers possible alternatives to formal standardised tests for reporting on language proficiency. It then describes a study that set out to discover students' and their future employers' views on the introduction of such an exit mechanism. The paper concludes by suggesting how a valid and reliable reporting mechanism could be developed for students in Hong Kong, by drawing on the current work being done by the Council of Europe's European Portfolio Project.

Introduction

Since 1997, when sovereignty over Hong Kong changed from British rule to that of the People's Republic of China, Hong Kong has to some extent struggled to establish its own identity. On the one hand is the desire to be divested of all trappings of the colonial past; on the other is the knowledge that Hong Kong's viability in the commercial world is largely dependent on its positioning as a knowledge-based, international business community. This latter outlook has been at the root of the concern, much discussed in the local media over the past two decades, that standards of English language proficiency in Hong Kong are declining.

A succession of Education Committee reports and Policy Addresses by the Governor (prior to 1997) and Chief Executive (following the return of Hong Kong to China) have led to the introduction of a number of schemes designed to strengthen language proficiency throughout the educational system. In the post-compulsory education sector these include, *inter alia*, generous grants to the eight tertiary institutions for language enhancement provision[i]; the development of a new benchmark test, the Language Proficiency Assessment

of Teachers (LPAT), to assess the language ability of English teachers in an attempt to increase the effectiveness of the teaching profession[ii]; and the instigation of a Workplace English Campaign in which English benchmark levels are aligned with the business English tests administered by four internationally recognised examination bodies, thereby offering employers and employees a range of choices as to which test to take.[iii] In order to address the widely held perception that the English language proficiency of university graduates is inadequate to meet the needs of a modern, multi-national workforce, it has recently been proposed that an 'exit-test' should be taken by all students as a condition of graduation.[iv]

In this paper we argue that introducing an exit-test as a sole measure of language proficiency would merely provide a 'quick-fix', short-term political solution, which is incompatible with published educational objectives for Hong Kong. We begin by outlining some of the background factors relevant to the current situation. Next, we argue that it is possible to accommodate the needs of employers whilst at the same time fostering an educational system that could nurture the development of life-long learning, but that to do so requires a willingness to challenge the status quo and develop more imaginative reporting procedures. We then present results of stakeholder studies carried out to elicit the views of students and employers from a range of companies. The paper concludes with suggestions for the development of a reporting mechanism, which will, we hope, be perceived as valid, reliable, useful and fair to all concerned.

Background

Despite the willingness of the government to tackle what is perceived to be a growing problem through the introduction of the measures mentioned above, business leaders still regularly express concern that students graduating from Hong Kong's tertiary institutions do not possess the requisite language skills to perform adequately in the modern workplace. The major local English language newspaper, the South China Morning Post, regularly publishes letters to the Editor and articles written by academics and businessmen lamenting the 'fact' (often with little more than 'anecdotal evidence') that students' English proficiency is not adequate to meet job demands. However, the general consensus amongst business leaders appears to be that it is not within the remit of business to provide training in general language skills for future employees (Lee and Lam 1994) and that there is a 'need for school leavers and graduates to have a good command of English to enter the business, professional and service sectors' (Education Commission 1995: 4).

In order to 'help employers choose employees whose English is at the right standard for the company's needs' (Hamp-Lyons 1999: 139), funding was awarded to the Hong Kong Polytechnic University in 1994 for the

development and trialling of test batteries in English and Chinese. The first full official administration of the resulting Graduating Students' Language Proficiency Assessment in English (GSLPA-English) took place in the 1999–2000 academic year. The content of the GSLPA-English is explicitly geared towards the types of professional communication that it is believed new graduates will face in their careers and is not restricted to the content of any single course at any one institution. There is no pass/fail element to the test as such; candidates receive a certificate that simply provides a description of their performance for both written and spoken English[v].

Tertiary institutions in Hong Kong have, however, strongly resisted the introduction of this test for a number of reasons, not the least of which is the fact that currently there are no degrees awarded in Hong Kong where common performance mechanisms are required across institutions. Another major concern of language educators within the institutions is the impact that the imposition of an exit-test would have on the existing curriculum. Although Alderson and Wall (1993) and Wall (1996; 1997) argue that there is little evidence either in general education or language education to support the notion that tests actually influence teaching, it is generally believed, in Hong Kong as elsewhere, that high-stakes testing programs strongly influence curriculum and instruction to the extent that the content of the curriculum is narrowed to reflect the content of a test. The relationship between teaching, testing and learning is therefore considered to be one of curricular alignment (Madaus 1988; Smith 1991; Shepard 1993). Madaus (1988) argues that the higher the stakes of a test, the more will be the impact, or washback on the curriculum, in that past exam papers will become the *de facto* curriculum and teachers will adjust their teaching to match the content of the exam questions (see also Shohamy 1993; 1998 and 2001). If, as has been suggested, students were in future required to achieve a certain degree of language competence as a condition of graduation, or for gaining entry into high-status professions, then the stakes would be very high indeed.

Messick (1996: 241) defines washback as 'the extent to which the use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit learning'. In the past few years revisions to many of Hong Kong's public examinations have been made in a deliberate attempt to provoke changes in teaching and learning although, as Hamp-Lyons (1999) notes, while there may have been discernible changes in the content of what is taught, there is little evidence that assessment change or innovation has led to changes in actual teaching practice. There remains, nevertheless, a concern among Hong Kong's tertiary educators that achieving a high score on the test would become the major focus of university study, to the detriment of other important skills. Whatever the resistance of educators and institutions, it is clear that in the near future some mechanism for reporting on students' language ability on graduation will be adopted in Hong Kong.

Challenging the status quo

One of the aims of any civilised society in the twenty-first century should be to encourage life-long learning with the concomitant upgrading of the skills of all its citizens. In Hong Kong, the aim is to develop to the greatest extent possible, a tri- or even quadri-lingual, bi-literate society in which heritage and target languages of professional communication can be valued equally. This means recognising the commensurability of spoken Chinese, the mother tongue (if different), standard Mandarin and English and written standard Chinese and English. In addition, skills such as problem solving, and critical thinking are being actively fostered. According to a recent Education Commission report, the main purpose of the education reform currently under way in Hong Kong is 'to give students more room and flexibility to organise and take charge of their own learning' (2000: 36), further warning: 'There should not be, at any stage of education, dead-end screening that blocks further learning opportunities' (ibid).

Considering the magnitude of the Education Commission's aspirations, it is highly unlikely that the simple introduction of an exit test would enable them to be achieved. Much more wide-ranging and imaginative reform is required and, because the goals for Hong Kong's society appear to have much in common with those expressed by the Council of Europe (Schärer 2000) it may be that Hong Kong could derive considerable benefit from the experiences of the European Language Portfolio project. The advantages of portfolio assessment have been thoroughly described in the literature (see, for example, Moya and O'Malley 1994; Daiker *et al*. 1996; Hamp-Lyons and Condon 2000). Genesee and Upshur (1996: 99) define a portfolio as 'a purposeful collection of students' work that demonstrates to students and others their efforts, progress and achievement in given areas'. Using terms that resonate strongly with the main purpose of the educational reforms outlined in the Education Commission's (2000) report, they describe some of the benefits of portfolio assessment as promoting *inter alia*:

- Student involvement in assessment
- Responsibility for self-assessment
- Student ownership of and responsibility for their own learning
- Excitement about learning
- Students' ability to think critically (op cit: 100).

Agreeing with many of Genesee and Upshur's comments, Berry and Lewkowicz (2000) also point out that portfolio assessment is on-going and can demonstrate progress over time as well as achievement. They emphasise that if learners were encouraged to compile a portfolio of their work, the contents need not be restricted to work produced specifically for the language classroom, arguing that a portfolio could, and perhaps should, include work that students have produced for their own faculty or even as part of their extra-

curricular activities. This has the potential to provide positive washback on both the English language classroom and the English-medium subject-content classroom. Although much of portfolio assessment in the past has been concerned with the assessment of writing, Genesee and Upshur (1996), Douglas (2000) and Berry and Lewkowicz (2000) all argue that portfolios can also be used effectively to record achievement in spoken language. Portfolio assessment could then be orientated more towards learners' individual needs and could provide a much more comprehensive evaluation of underlying and wide-ranging abilities than is possible with a single test outcome. Learners could also be encouraged to revise and work on submissions reflecting authentic contexts so that each individual portfolio represented a record of achievements they were proud to show to potential employers, in much the same way as art and graphics students have a portfolio to show at the end of their studies.

As with all methods of assessment, portfolio assessment is not without some disadvantages. In a study carried out by Berry and Lewkowicz (2000), students[1] identified the following as constituting potential problems:

- Portfolios are time-consuming to compile and could involve a great deal of extra work over and above normal faculty and language class requirements
- There is a danger of plagiarism and the possibility of far too much teacher input in revising work for the end product to be considered a true reflection of the learner's ability
- The end product is subject to considerable variation in responsiveness and also to problems of interpretability by the intended recipients
- Portfolios are difficult to assess in a standardised way and would require considerable rater training to be rated fairly, effectively, reliably and validly.

These last two points are also addressed by Douglas (2000: 244) who, citing Moya and O'Malley (1994) and Brown and Hudson (1998), draws attention to the need for demonstrating the validity and interpretability of portfolio assessment.

None of these problems is insuperable, however, and we believe that the pedagogic advantages of portfolio assessment significantly outweigh the

1 Berry and Lewkowicz (2000) report the results of a pilot survey carried out at the University of Hong Kong in April 2000 to solicit views of undergraduate students about the need for and desirability of introducing compulsory language assessment prior to graduation. 1418 students took part in the study, which was designed to elicit students' views on the following issues: 1) whether at graduation students should be required to take a language test to demonstrate their proficiency in English; 2) what authority should be responsible for administering the test if a test were required; and 3) whether they considered a portfolio to be a fair and acceptable alternative to a language test.

administrative disadvantages. As with the European portfolio, we recognise that there could be a variety of portfolio models; we also take heed of Lucas' warning that if portfolios become standardised and contain equivalent tasks they 'will be just as likely as other standardised tests to limit learning by restricting curriculum to what is most easily and economically measured' (Lucas 1992: 9, quoted in Yancey and Weiser 1997: 13). Nevertheless, for purposes of transparency, coherence, reliability and comparability in the reporting of linguistic and socio-cultural competencies, we believe there would need to be a common core and a uniform way of reporting information.

Stakeholder Surveys

Before a new system could be suggested, it was necessary to determine whether Hong Kong was ready to accept a challenge to current beliefs and practices. To ascertain this and to investigate the types of change that would be acceptable, we carried out two surveys of the primary stakeholders likely to be affected by any change in policy, namely students and employers. These are described below.

Student Survey

The purpose of this survey was to determine students' perceptions of the most effective way to report their English language ability on graduating from university. To tap the opinions of as many students as possible across a range of tertiary institutions, we decided to use a questionnaire. This questionnaire was based on a pilot questionnaire developed at the University of Hong Kong (HKU) in the previous year (for details see Berry and Lewkowicz 2000). The revised questionnaire that was again piloted, this time on a small group of HKU students, was divided into five sections, eliciting: demographic data; alternative elements which could be included in a portfolio; the usefulness of different forms of test that could be used; optimal ways of providing future employers with useful and accurate information about applicants' language proficiency; and a final section inviting additional comments (see Appendix 1).

Responses were received from 1600+ students from seven of the eight tertiary institutions in Hong Kong. The majority were from first- or second-year students, 69% and 26% respectively, as they were the most likely to be affected by any change in policy. They came from a range of disciplines from engineering and science to the arts, law and the humanities, and for each targeted discipline there were respondents from at least two institutions. Although the number of respondents varied across institutions from 40 to 400+, the similarity of responses suggests that there is considerable agreement among students as to how their language ability should be reported.

Even though most of the students (71%) had been learning English for 16 or more years, and the majority (77%) had been through an English-medium secondary education, they reported that their exposure to English outside the formal classroom setting was limited (see Table 1). Despite this, and the recognition of many respondents (65%) that their English had not improved at university, they were realistic in recognising that they would be required to demonstrate their English ability on graduation and they were co-operative in completing the questionnaire. None took the opportunity of the additional comments section to say that their level of English was irrelevant or that none of the ways suggested for reporting their level was appropriate.

Table 1 Frequency of opportunities to speak English

	Never		Occasionally		Often	
	No.	Per cent	No.	Per cent	No.	Per cent
With native speakers	275	17.1	828	51.4	435	27.0
With family and/or friends	803	49.8	615	38.2	108	6.7
With speakers of other languages	407	25.3	860	53.4	260	16.1

When asked to rank-order three possible options for reporting their English ability to employers, namely using a portfolio which includes a recent test score, a portfolio with no test score, or a test score on its own, the first option – that is, portfolio plus test score – appeared to be the one most favoured (see Table 2). This suggests that the majority of the students are pragmatic and recognise that for any report of their language ability to be informative, it should be comprehensive. Had they selected the most expedient solution, they would, most probably, have opted simply for a test score.

Table 2 Students' rankings of three options for demonstrating language proficiency

	Test only		Portfolio only		Portfolio + test	
Rank	No.	Per cent	No.	Per cent	No.	Per cent
1st	142	8.8	381	23.6	1049	65.1
2nd	410	25.5	761	47.2	449	27.9
3rd	1035	64.2	446	27.7	92	5.7

If a test were to be mandated, it would appear that students would prefer to take one that is internationally recognised. Given the choice between an international test such as, for example, IELTS, a Hong Kong-wide test developed locally for all Hong Kong students and a university specific test, the

majority selected an international test as their first choice and the Hong Kong-wide test as their second choice, the frequencies of response being 939 (58%) and 861 (53%) respectively (see Table 3).

Table 3 Students' rankings of three types of test

Rank	University Specific		Hong Kong Wide		International	
	No.	Per cent	No.	Per cent	No.	Per cent
1st	192	11.9	471	29.2	939	58.3
2nd	417	25.9	861	53.4	348	21.6
3rd	991	61.5	267	16.6	312	19.4

Giving reasons for preferring an international test, students commented that such tests are more widely recognised, thereby allowing Hong Kong students to compete and be compared with international students. Students opting for the Hong Kong-wide test saw the fact that the test was specific to Hong Kong as an advantage and thought such a test would be more informative for employers, while those who preferred the university-specific test considered that such a test would be the easiest to organise.

Students were also asked to select from a list drawn up on the basis of 'items' proposed by students in the HKU pilot study (Berry and Lewkowicz 2000) what should be included in a language portfolio if one were to be required as part of their final assessment. Not surprisingly, perhaps, there was considerable variability in responses in this section. While there was some agreement that the portfolio should include a record of the students' work, with the majority of respondents considering the following elements as integral: a résumé (74%), language scores/grades achieved at university (52%), and their '*Use of English*'[vi] score (49%), there was less agreement about the samples of work that should be presented. Although the majority (56%) considered a written introduction of themselves should be included by everyone, most respondents thought other samples of work should be left to the discretion of individual students. This, to a large extent, reflects the underlying philosophy of portfolio assessment, which stresses the importance of giving students choices and allowing them to demonstrate their strengths. However, the comparatively large proportion of respondents (about one-third) who rejected as unnecessary the inclusion of any form of oral performance, suggests that there would be a need to inform students of the value of demonstrating their oral abilities if portfolio assessment were to be accepted by employers (see below).

Survey of Employers

The aim of this survey was to determine whether potential employers, from both large and small companies, would accept portfolio assessment. Representatives from the Human Resources Departments of 12 companies, four multi-nationals, four public-listed companies and four small-medium enterprises were invited to participate in in-depth, structured interviews at their convenience. Each interview lasted 45–60 minutes, with the interviewers taking notes during the interview; where possible, the interviews also were audio-recorded. The interviews, which followed the organising framework of the student questionnaires, centred on the companies' selection procedures for new graduates, their language requirements, and their attitudes towards exit tests and portfolio assessment (see Appendix 2). Here we focus on the results relating to portfolio assessment and tests.

During the interviews it became clear that concepts such as portfolio assessment, which are well known within educational circles, were not at all familiar to the business community. It was therefore necessary to provide interviewees with information about portfolio assessment and show examples of portfolios such as those being developed for use in Europe. Despite their initial lack of familiarity with alternative methods of assessment, once the human resources representatives saw what this would involve, the majority (7 of 12) ranked the option of portfolio including a test score as the best. However, they added a proviso that employers should not be responsible for going through applicants' portfolios but that each portfolio should contain a one-page summary verified by an assessment professional.

The company representatives went to some length to explain that their main purpose in seeing a language assessment at graduation was not simply to help them with their selection procedures, as they insisted that they would continue to use their own selection procedures whether or not a language assessment was introduced. They were, instead, adamant that there was a need for a mechanism that would ensure the improved language abilities of those graduating from Hong Kong universities. Complementing their reluctance to take responsibility for examining applicants' portfolios, they also stressed that the format of the assessment should be determined by those qualified in the field, emphasising that they did not feel it was their responsibility to say what should be included in a portfolio if one were to be introduced.

During the discussions it became apparent that employers were often looking for more than language skills; they also wanted the graduates to have improved socio-linguistic competence, which is, of course, very difficult to assess using traditional tests. In addition, they were looking for enhanced oral as well as written skills, preferably to the level of the many returning graduates who had studied outside Hong Kong.

Whereas the majority of employers were in agreement with the students

that a comprehensive means of reporting language ability would be most beneficial, they were in less agreement as to the type of test that would be most appropriate. Of the 12 representatives, seven ranked a Hong Kong-wide test as their preferred option, while three, all from multi-national companies, considered an international test would be the best. The remaining interviewee commented that either of the above would be acceptable. None opted for individual universities developing their own tests.

Conclusions

Perhaps one of the main limitations of the student survey is that even though some of the respondents had undoubtedly experienced compiling a language portfolio during their time at university, there is no guarantee that all were fully aware of what this would entail. Despite this, it appears that students would favour providing comprehensive evidence of their language abilities and that they are ready and able to participate in any discussions of future assessment requirements. Employers also seem prepared to accept changes to current assessment practices and so a way forward would be for Hong Kong to learn from the seminal work being done in Europe, but to accept that any system introduced would need to be modified for the particular circumstances of Hong Kong. This would inevitably take time, especially as an acceptable reporting mechanism would have to be developed. It would also be necessary to raise consciousness among the different stakeholders. Our survey of Human Resources representatives, though restricted, showed that most employers need information about portfolio assessment if it is to be introduced. It also showed that there might be a mismatch between employers' expectations as to the language abilities they consider graduates need to enhance, and those abilities that students deem it necessary to demonstrate.

These are not, however, insurmountable problems and if portfolio assessment were embraced it would go some way towards ensuring that the system implemented matched Hong Kong's educational objectives. Furthermore, if such a system were developed in conjunction with a structured and rigorous program of research, Hong Kong could end up with an assessment mechanism that was not only valid and reliable but was also useful and fair to all.

Acknowledgments

This research was supported by C.R.C.G. grants 10202627 and 10203331 from the University of Hong Kong. We would like to thank our research assistant, Matthaus Li, for assistance with data input and analysis.

References

Alderson, J. C. and D. Wall. 1993. Does washback exist? *Applied Linguistics* 14: 2, 115–129.

Berry, V. and J. Lewkowicz. 2000. Exit tests: Is there an alternative? In V. Berry and J. Lewkowicz (eds.) *Assessment in Chinese Contexts: Special Edition of the Hong Kong Journal of Applied Linguistics*: 19–49.

Brown, J. D. and T. Hudson. 1998. The alternatives in language assessment. *TESOL Quarterly* 32: 4, 653–676.

Daiker, D., J. Sommers and G. Stygall. 1996. The pedagogical implications of a college placement portfolio. In E. White, W. Lutz and S. Kamusikiri (eds.) *Assessment of Writing* (pp. 257–270). New York: The Modern Language Association of America.

Douglas, D. 2000. *Assessing Languages for Specific Purposes.* Cambridge: Cambridge University Press.

Education Commission. 1988. *Education Report 3. The Structure of Tertiary Education and the Future of Private Schools*. Hong Kong: Government Printer.

Education Commission. 1995. *Education Report 6. Enhancing Language Proficiency: A Comprehensive Strategy*. Hong Kong: Government Printer.

Education Commission 2000. *Learning for Life; Learning through Life: Reform Proposal for the Education System in Hong Kong*. Hong Kong: Government Printer.

Falvey, P. and D. Coniam. 2000. Establishing writing benchmarks for primary and secondary teachers of English language in Hong Kong. In V. Berry and J. Lewkowicz (eds.) *Assessment in Chinese Contexts: Special Edition of the Hong Kong Journal of Applied Linguistics*: 128–159.

Genesee, F. and J. A. Upshur. 1996. *Classroom-Based Evaluation in Second Language Education*. Cambridge: Cambridge University Press.

Hamp-Lyons, L. 1999. Implications of the 'examination culture' for (English language) education in Hong Kong. In V. Crew, V. Berry and J. Hung (eds.) *Exploring Diversity in the Language Curriculum* (pp. 133–140). Hong Kong: The Hong Kong Institute of Education.

Hamp-Lyons, L. and W. Condon. 2000. *Assessing the Portfolio: Principles for Practice, Theory, Research*. Cresskill, NJ: Hampton Press.

Lee, N. and A. Lam. 1994. *Professional and Continuing Education in Hong Kong: Issues and Perspectives.* Hong Kong: Hong Kong University Press.

Lucas, C. 1992. Introduction: Writing portfolios-changes and challenges. In K. B. Yancey (ed.) *Portfolios in the Writing Classroom: An Introduction* (pp. 1–11). Urbana, Illinois: NCTE.

Madaus, G. F. 1988. The influence of testing on the curriculum. In L. N. Tanner (ed.) *Critical Issues in the Curriculum: 87th Yearbook of the National Society for the Study of Education, Part 1* (pp. 83–121). Chicago: University of Chicago Press.

Messick, S. 1996. Validity and washback in language testing. *Language Testing* 13: 3, 241–256.

Moya, S and J. M. O'Malley. 1994. A portfolio assessment model for ESL. *The Journal of Educational Issues for Minority Students*, 13: 13–36.

Schärer, R. 2000. *A European Language Portfolio: Third Progress Report*. The Council of Europe. Retrieved from: http://culture.coe.fr/lang/-Portfolio/eng/3rdreport.htm, December 2000.

Shepard, L. A. 1993. Evaluating test validity. *Review of Research in Education* 19: 405–50. Washington, DC: American Educational Research Association.

Shohamy, E. 1993. *The Power of Tests: The Impact of Language Tests on Teaching and Learning. NFLC Occasional Papers*. Washington DC: The National Foreign Language Center.

Shohamy, E. 1998. Language tests as *de facto* educational and linguistic policies. In: V. Berry and McNeill (eds.) *Policy and Practice in Language Education* (pp. 23–41). Hong Kong: Department of Curriculum Studies, The University of Hong Kong.

Shohamy, E. 2000. Using language tests for upgrading knowledge: The phenomenon, source and consequences. In V. Berry and J. Lewkowicz (eds.) *Assessment in Chinese Contexts: Special Edition of the Hong Kong Journal of Applied Linguistics*: 1–18.

Shohamy, E. 2001. *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. New York: Pearson Education Ltd.

Smith, M. L. 1991. Put to the test: The effects of external testing on teachers. *Educational Researcher* 20: 5, 8–11.

Wall, D. 1996. Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing* 13: 3, 334–354.

Wall, D. 1997. Impact and washback in language testing. In C. Clapham and D. Corson (eds.) *Encyclopedia of Language and Education Vol. 7: Language Testing and Assessment* (pp. 291–302). Dordrecht, Boston, London: Kluwer Academic Publishers.

Yancey, K. B. and I. Weiser. 1997. *Situating Portfolios. Four Perspectives.* Logan, Utah: Utah State University Press.

Appendix 1

Questionnaire

It is probable that students will be required to demonstrate their English language proficiency on graduating from university. Since there are a number of options for reporting proficiency levels, we are interested in your views as to which option you consider would best demonstrate your proficiency in English.

This booklet contains FOUR pages including the front and back. Please answer all questions, including those on the back page.

All responses to this questionnaire will be treated with the utmost confidentiality and used for research purposes only.

Thank you for your co-operation.

I. Personal Data:

1. Gender (Please circle as appropriate) Male Female
2. Language (s) spoken at home: (Please tick as appropriate)
   Cantonese: _____ Other Chinese dialects (including Putonghua): _____
   English: _____ Other (please specify): _____
3. Language of Secondary Education: (Please tick as appropriate)
   English (EMI): _____ Chinese (CMI): _____
   Other (please specify): _____
4. Please circle year of study at university: 1 2 3 4 5
5. University Degree/Major: _____
6. Number of years you have been studying English (please tick as appropriate):
   12 or less: _____ 13_____ 15: _____ 16 or more: _____
7. Languages other than English and Cantonese spoken, or studied at school or university (Please specify):
   _____
8. Please tick all the following places you have been to outside Hong Kong:
   PRC/Macau/Taiwan: _____ Rest of Asia: _____ UK:_____ Europe: _____
   Africa:_____ USA/Canada:_____ Australia/New Zealand: _____
9. When travelling outside Hong Kong, how often did you speak English to communicate with (Please tick as appropriate):

	Never	Occasionally	Often
Native speakers of English	__	__	__
Family members/friends	__	__	__
Speakers of languages other than English or Cantonese	__	__	__

10. Since starting university my level of English has (Please tick as appropriate):
    Improved:_____ Stayed the same: _____ Declined: _____

II. Portfolios

Portfolio assessment is becoming increasingly popular around the world since it represents an opportunity for students to provide extended evidence of what they can do in one or more languages and in a range of areas. What is included in a portfolio may be pre-specified or left to individuals to select, or a combination of both. It may be collected over an extended period of time.

Below is a range of items which could be included in a graduating student's language portfolio. Consider each element and indicate whether it should be a compulsory or optional element of the portfolio, or whether it should not be included at all.

Alternative elements which could be included in a Portfolio

	Compulsory	Optional	Not to be included
Résumé (C.V.)			
Video introducing oneself			
Written introduction of oneself			
Examples of business correspondence			
Examples of project work			
Written commentary on a current affairs issue			
Academic writing (e.g. essays, reports, etc. marked by faculty)			
Academic writing (e.g. essays, reports, etc. produced for English enhancement classes)			
Writing done under timed conditions in class			
Video of a formal oral presentation			
Video of other oral skills, e.g. group discussions, role plays, etc.			
Self-assessment of your language skills			
Peer assessment of your language skills			
Teacher assessment of your language skills			
HKEA 'Use of English' grade			
Language scores/grades achieved at university			
Record of any language-related work or formal courses you have taken			
Record of language experiences outside Hong Kong (e.g. on holiday or travelling)			

Can you think of other examples of work that you would like included in a portfolio. Please specify and indicate whether each should be optional or compulsory:

Examples	Compulsory	Optional	Not to be included

III. Tests

Another way of showing your language proficiency is by simply providing a test score. Below are three alternatives, all of which are being considered. Please rank these according to which you think would be the most useful to you and to your future employers (1 = most useful and 3 = least useful)

University-specific test (each institution sets and marks its own tests) ___
Hong Kong-wide test (one test for all graduating students in Hong Kong) ___
International test (a test developed outside Hong Kong and
widely recognised throughout the world, e.g. IELTS; TOEFL) ___

Please give reasons for your FIRST CHOICE.

_____
_____
_____

IV. Optimal Solution

Which of the following options do you think would best provide your future employers with useful and accurate information about your English language proficiency? Please rank the options (1 = best and 3 = worst)

Language Portfolio (only): ___

Language Portfolio which includes a test score for a recently taken test: ___

Recent test score (only): ___

Please give reasons for your FIRST CHOICE.

_____
_____
_____

V. Additional Comments:

Please add any further comments you have about demonstrating your English language proficiency on graduation. You may continue your comments on the first page of this questionnaire.

_____
_____
_____

Thank You

Appendix 2

Company Interview Schedule

It is highly likely that in the near future university students will be required to demonstrate their English language proficiency at graduation. Since there are a number of options available for reporting proficiency levels, we are extremely interested in your views as to which option or options you consider would best demonstrate students' proficiency in English. We would also like you to consider the various potential components for each of the options and suggest which would provide the most useful information for your recruitment needs.

All responses will be treated with the utmost confidentiality and used for research purposes only.

Thank you for your co-operation.

Name of company: _____

Type of company: Multinational: _____ PLC: _____ SME: _____

Name of person interviewed: _____

Job title: _____

Date: _____

I. Company Data

1. On average, approximately how many new graduates do you recruit each year?

_____

_____

2. What sorts of entry level positions are available?

_____

_____

3. Do you test potential recruits for language skills? Yes No

4. If you answered yes to question 3, what form does the test take?

_____

_____

5. What do you perceive as typical language problems in potential recruits?

_____

_____

6. Can you tell us what you know about the government's proposal for graduating students' exit assessment?

_____

_____

7. Are you in favour of it? Yes No

8. Are you aware of the possibility of using a language portfolio for assessment purposes? Yes No

9. If you were to use a portfolio to assess language skills, would you require a cover page providing summary information? Yes No

10. Can you suggest what information would crucially be contained on a cover page, if required?

_____

_____

II. Portfolios

Portfolio assessment is becoming increasingly popular around the world since it represents an opportunity for students to provide extended evidence of what they can do in one or more languages and in a range of areas. What is included in a portfolio may be pre-specified or left to individuals to select, or a combination of both. It may be collected over an extended period of time.

Below is a range of items which could be included in a graduating student's language portfolio. Please consider each element and indicate whether it should be a compulsory or optional element of the portfolio, or whether it should not be included at all.

Alternative elements which could be included in a Portfolio

	Compulsory	Optional	Not to be included
Résumé (C.V.)			
Video introducing the student			
Written introduction of the student			
Examples of business correspondence			
Examples of project work			
Written commentary on a current affairs issue			
Academic writing (e.g. essays, reports, etc. marked by faculty)			
Academic writing (e.g. essays, reports, etc. produced for English enhancement classes)			
Writing done under timed conditions in class			
Video of a formal oral presentation			
Video of other oral skills, e.g. group discussions, role plays, etc.			
Self-assessment of students' language skills			
Peer assessment of students' language skills			
Teacher assessment of students' language skills			
HKEA 'Use of English' grade			
Language scores/grades achieved at university			
Record of any language-related work or formal courses taken			
Record of language experiences outside Hong Kong (e.g. on holiday or travelling)			

Can you think of other examples of work that you would like included in a portfolio. Please specify and indicate whether each should be optional or compulsory:

Examples	Compulsory	Optional	Not to be included

III. Tests

Another way of showing students' language proficiency is by simply providing a test score. Below are three alternatives, all of which are being considered. Please rank these according to which you think would provide the most useful information to you for recruitment purposes (1 = most useful and 3 = least useful)

University-specific test (each institution sets and marks its own tests) ___

Hong Kong-wide test (one test for all graduating students in Hong Kong) ___

International test (a test developed outside Hong Kong and widely recognised throughout the world, e.g. IELTS; TOEFL) ___

Please give reasons for your FIRST CHOICE.

_____

_____

_____

IV. Optimal Solution

Which of the following options do you think would best provide you with useful and accurate information about students' English language proficiency? Please rank the options (1 = best and 3 = worst)

Language Portfolio (excluding a test score): ___

Language Portfolio which includes a test score for a recently taken test: ___

Recent test score only: ___

Please give reasons for your FIRST CHOICE.

_____

_____

_____

V. Additional Comments

Please add any further comments you have about the issue of students demonstrating their English language proficiency on graduation.

_____

_____

_____

Thank You

Endnotes

i Language Enhancement Grants were initiated in response to a recommendation in the Education Commission's third report (ECR3), June 1988: 85. The eight institutions are: City University of Hong Kong, Hong Kong Baptist University, Hong Kong Institute of Education, Lingnan University, The Chinese University of Hong Kong, The Hong Kong Polytechnic University, The Hong Kong University of Science and Technology and The University of Hong Kong.

ii Falvey and Coniam, 2000, discuss aspects of the development of the LPAT; Shohamy, 2000 offers an alternative perspective on this initiative

iii The four tests are the Business Language Testing Service (BULATS) see http://www.ucles.org.uk, English Language Skills Assessment (ELSA) see http://www.lccieb.com/LCCI/ Home/ELSA.asp/, Test of English for International Co-operation (TOEIC) see http://www.toeic.com/ and Pitman's Tests of English for Speakers of Other Languages (EOS1; EOS2) see http://www.city-and-guilds.com.hk/pq/wpestart.htm

iv The Chief Executive of the HKSAR, in his first Policy Address on October 8 1997, announced that he hoped universities would consider introducing a requirement for all graduating students to take proficiency tests in English and Chinese. See http://www.info.gov.hk/ce/speech/cesp.htm, Section E, Paragraph 93.

v For a complete description of the current version of the GSLPA-English, see http://www.engl.polyu.edu.hk/ACLAR/projects.htm#GSLPAdevt

vi This is a public examination run by the Hong Kong Examinations Authority, which students have to pass to enter university. See http://www.hkea.edu.hk

5 Validating questionnaires to examine personal factors in L2 test performance

James E. Purpura
Teachers College, Columbia University

Introduction

It has long been established that individual learner characteristics may contribute differentially to a student's ability to acquire a second language (Skehan 1989, 1998). Similarly, it has been shown that certain test-taker characteristics, apart from that of communicative language ability, may also influence the degree to which test takers are able to perform optimally on language tests (Bachman 1990). Witness to this is a large body of research demonstrating that, aside from language knowledge, the personal attributes of test takers have a significant effect on test-score variation. Most of this research has focused on the relationship between test takers' demographic characteristics and their performance on tests. Such studies have examined performance with relation to age (e.g. Farhady 1982; Spurling and Illyin 1985; Zeidner 1987), gender (e.g. Farhady 1982; Kunnan 1990; Ryan and Bachman 1990; Sunderland 1995; Zeidner 1987), cultural background (e.g. Brière 1968; Farhady 1979; Zeidner 1986, 1987), and language background (e.g. Alderman and Holland 1981; Brown 1999; Brown and Iwashita 1996; Chen and Henning 1985; Elder 1995; Farhady 1982; Ginther and Grant 1997; Kunnan 1990, 1995; Oltman *et al*. 1988; Ryan and Bachman 1990; Swinton and Powers 1980).

A second body of research has looked at the relationship between test takers' topical knowledge and their performance on language tests. Many of these studies have investigated performance in relation to test takers' academic background or their prior knowledge (e.g. Alderson and Urquhart 1985; Clapham 1993, 1996; Fox *et al*. 1997; Ginther and Grant 1997; Jensen and Hansen 1995; Tedik 1990).

A third set of studies has examined the socio-psychological and strategic characteristics of test takers and their performance on tests. These studies have examined performance in relation to cognitive styles such as field dependence

and independence (e.g. Chapelle 1988; Hansen and Stansfield 1984, Stansfield and Hansen 1983), attitudes toward language learning (e.g. Clément and Kruidenier 1985; Gardner 1985, 1988; Zeidner and Bensoussan 1988); motivation and the degree to which test takers are willing to devote time and effort to language learning (e.g. Clément and Kruidenier 1985; Dörnyei 1990; Dörnyei and Schmidt 2001; Gardner 1985, 1988; Gardner and Lambert 1972; Kunnan 1995), level of anxiety (e.g. Brown *et al*. 1996; Bensoussan and Zeidner 1989), and the test takers' capacity to use cognitive and metacognitive strategies effectively (Anderson *et al*. 1991; Purpura 1999; Vogely 1995). These socio-psychological and strategic factors, alone or in combination with other personal attributes, may have a significant impact on test scores, suggesting that language knowledge may be a necessary, but in fact not a sufficient, condition for 'good' language test performance.

Given the potential role of these factors in second-language learning and assessment, researchers must continue to investigate the nature of learner characteristics and their potential effects on learning outcomes. They must also examine how these attributes interact with each other, and how their simultaneous effect contributes to test-score variation; otherwise, the very constructs we wish to measure may be masked.

Prior to examining these relationships, however, valid and reliable instruments designed to measure these attributes must be developed. One well established method for assessing test-taker characteristics is the questionnaire. Questionnaires allow for a high degree of control over the probes; they can be easily designed to measure multiple constructs simultaneously; they can be administered to large groups of examinees; they lend themselves to statistical analysis; and they reveal systematic patterns of behaviour in large amounts of data that might otherwise have gone unnoticed. However, questionnaires are notoriously sensitive to small differences in wording (Allan 1995); they often show cross-measurement of content, producing substantial redundancy and correlated measurement error (Byrne 1998; Purpura 1998, 1999); and they produce over- or underestimates of the data. Given these problems, it is important that the construct validity of questionnaires be thoroughly investigated prior to their use in research or their application to learning, and validation efforts need to be substantially and methodologically rigorous. Otherwise, the inferences drawn from the use of these instruments may be unfounded and misleading.

While the development and validation of instruments purporting to measure these personal attributes are a critical first step in examining the relationships between personal factors and performance, most questionnaires currently being used have not been submitted to such rigorous validation procedures. In fact, most researchers report no more than an assessment of the questionnaire's internal consistency reliability. However, an increasing number of researchers (e.g. Gardner 1985b; Oxford *et al*. 1987; Purpura 1997,

1998) carried these analyses a step further by examining the underlying factor structure of their questionnaires by means of exploratory factor analysis. No study in our field, however, has used a confirmatory approach to examining the factor structure of items in a questionnaire. In the current study, item-level structural equation modelling (SEM) has been used as a means of examining the underlying psychometric characteristics of questionnaire surveys.

The current paper presents the preliminary findings of an on-going study aimed at examining the construct validity of a battery of questionnaires designed to measure selected socio-psychological and strategic background characteristics of test takers. I will first describe these measures and the theoretical constructs underlying their construction. I will then discuss the process used to examine the construct validity of these instruments, using item-level structural equation modelling, and how these validation efforts have informed decisions about tailoring the instruments prior to their computerisation and use in research and learning contexts.

Background of the language learning questionnaires

The Language Learning Questionnaires (LLQs) were originally developed by Bachman, Cushing and Purpura (1993) in collaboration with the EFL division of UCLES. The goal was to provide a battery of questionnaires designed to measure selected background characteristics of the Cambridge EFL candidature and to examine how these characteristics might impact on performance in the Cambridge exams. The original bank of questionnaires concentrated on two sets of factors: selected socio-psychological factors and strategic factors. The socio-psychological questionnaire battery was designed to measure attitudes, motivation, effort and anxiety, while the strategic factor questionnaire battery was intended to measure cognitive, metacognitive and communication strategies (not discussed in this study), as seen in Table 1.

The development of the socio-psychological factor questionnaire battery was influenced by theories of motivation both in second language learning and in psychology. Initially, we were drawn to the work of Gardner (1979, 1985a) on second-language learning. In this work, the notion of 'attitude' was seen as an underlying and predisposing orientation towards motivational behaviours. Orientations, such as the commonly cited integrative or instrumental orientations, were considered as clusters of reasons for L2 study. Gardner (1985a) defined motivation as 'the extent to which an individual works or strives to learn the language because of a desire to do so and the satisfaction experienced in this activity' (p. 10). This definition involved three components: (1) motivational intensity or effort expended to learn the language, (2) a desire to learn the language, and (3) a positive attitude towards

Table 1
Taxonomy of the original language learning questionnaires

Socio-psychological factors	No. of Items
A. Attitudes Questionnaire	**41**
• Attitudes toward English speakers	13
• Attitudes toward learning English	8
• Interest in foreign languages	8
• Perception of task difficulty	12
B. Motivation Questionnaire	**57**
• Integrative motivation	17
• Instrumental motivation	12
• Achievement motivation – general learning	4
• Achievement motivation – language learning	11
• Achievement motivation – general testing	7
• Achievement motivation – language testing	6
C. Effort Questionnaire	**11**
D. Anxiety Questionnaire	**30**
• Class anxiety	9
• Language anxiety	9
• Test anxiety	12
SUB-TOTAL	**139**
Strategic factors	**No. of items**
A. Cognitive Strategies	**34**
• Clarifying/verifying	2
• Inferencing	2
• Summarising	2
• Analysing inductively	3
• Associating	4
• Linking with prior knowledge	4
• Repeating/rehearsing	5
• Applying rules	3
• Practising naturalistically	5
• Transferring from L1 to L2	4
B. Metacognitive Strategies	**21**
• Assessing the situation (planning)	6
• Monitoring	4
• Evaluating	5
• Self-testing	6
SUB-TOTAL	**55**
TOTAL	**194**

learning the language. In short, a motivated learner was said to display characteristics of all three components.

Probably the most compelling and the most frequently researched part of Gardner's (1985a) theory was his notion of integrative motive. This was defined as the 'motivation to learn a L2 because of positive feelings towards the community that speaks that language' (pp. 82–83). This construct was defined theoretically in terms of (1) integrativeness, or the willingness to integrate into the L2 community, (2) attitudes towards the learning situation and (3) motivation, or effort and desire to learn. Based on this theory, Gardner and Smythe (1975, 1981) developed the Attitude/Motivation Test Battery (AMTB), which has been until now 'the only published, standardised test of L2 motivation' (Dörnyei 2001: 52). This battery of questionnaires was designed to measure: integrativeness, attitudes toward the learning situation, motivation, instrumental orientation and language anxiety. Gardner's work served as a major influence in the design of the current battery of Language Learning Questionnaires.

The development of the socio-psychological factor questionnaires was also influenced by theories of attitudes and motivation in cognitive and educational psychology. For example, following the work of Eiser (1986), Hovland and Rosenberg (1960), and Brecker (1984), attitudes were viewed in terms of a three-component model involving (1) cognition, or statements of opinions and beliefs, (2) affect, or evaluative statements of feelings and preferences, and (3) behaviour, or statements of overt action or intent. As a result, each questionnaire contained items from these three perspectives. To illustrate, a cognitive response item eliciting beliefs and opinions might be: *This language is hard to remember*; an affective response item eliciting feelings and preferences might be: *This language is fun to learn*; and a conative response item eliciting readiness for action might be: *I study this language as often as I can*. This work was theoretically similar to that of Gardner and it helped us focus the types of questions to be included in the questionnaires.

The socio-psychological questionnaire battery was also influenced by our understanding of attribution theory as articulated by Weiner (1979, 1992). Attribution theory seeks to explain the causal determinants of past successes and failures because it is these factors that influence people's future achievement needs and motivations. Factors influencing achievement outcomes relate to three dimensions: (1) locus, or the degree to which success or failure is a result of the person or the situation, (2) stability, or the degree to which success or failure can be changed, and (3) controllability, or the degree to which success or failure can be controlled. According to Graham (1994), some of the attributions that account for success in school environments include: ability, effort, task difficulty, luck, mood, family background, help or hindrance from others. Of these, Weiner (1992) found that ability and effort had the greatest impact on learning. Further research has

also shown that success can also be attributed to the learner's perception of the task as being easy, while failure stems from the perception that the task demands appear unreasonable (McCombs 1991).

Based on this research, the Attitudes Questionnaire in the current study included four scales: attitudes toward speakers of English, attitudes towards learning a foreign language, and interest in learning a foreign language. Finally, because we felt that test takers' perception of the language as being difficult to learn might be an important factor, we included 'perception of task difficulty' as one of the scales.

The design of the Motivation Questionnaire was also rooted in Gardner's AMTB. In this case, we included instrumental and integrative motivation as scales in the current instrument. We were equally influenced by Weiner's (1979) notion of 'effort' as having motivational consequences resulting in success. According to Weiner (1979), success was a result of hard work, while failure was due to a lack of effort. Consequently, those who believe they have some degree of control over their success seem to exert more effort in pursuit of their goals. As achievement motivation and effort seemed to be potentially important factors in language learning we included these scales in the questionnaires. 'Achievement motivation' refers to beliefs and opinions about one's ability to achieve, while 'effort' refers to the concrete actions a learner is willing to do to achieve.

The final socio-psychological factor questionnaire was designed to measure 'anxiety', a condition which may undermine language learning or test performance. The AMTB defined anxiety in terms of the language class. However, FCE candidates may also experience anxiety associated with using the language in real-world communicative situations, where fears may surface as a result of lack of adequate linguistic control or lack of familiarity with the norms and expectations of the target culture. Also, the FCE candidates may experience anxiety related to taking language tests. As a result, the Anxiety Questionnaire sought to measure three types of anxiety: language class anxiety, language anxiety and test anxiety, as seen in Table 1.

The development of the strategic factors questionnaire battery was rooted in Gagné, Yekovich and Yekovich's (1993) model of human information processing and was influenced by several second-language strategy researchers. As the development of the cognitive and metacognitive strategy questionnaires in the LLQs is well documented in Purpura (1999), I will not duplicate that discussion here. Briefly, the Cognitive Strategy Questionnaire was designed to measure a number of comprehending, memory and retrieval strategies, while the Metacognitive Strategy Questionnaire aimed to measure a set of appraisal strategies such as monitoring and evaluating. For reasons of space, I will also not discuss the communication strategies questionnaire.

Once the questionnaires outlined in Table 1 were developed, they were piloted with students around the world (see Bachman, Cushing and Purpura

1993). They were then submitted to a series of reliability analyses, which allowed us to reduce the number of items and increase the homogeneity of the scales. In some cases, scales were dropped and others combined. Meanwhile, approximately 50 students were interviewed in relation to the questionnaires they took and the responses they provided. This allowed us to remove ambiguous items and simplify the wording of items, prior to the use of the questionnaires in the current study.

The current study

Purpose of the validation project

The purpose of the LLQ validation project was to examine the construct validity of the bank of language learning questionnaires so that they could be used to measure the background characteristics of the Cambridge EFL test takers. Once the psychometric characteristics of the questionnaires were known, the intent was to create a long (research) version and a short (classroom) version of the questionnaires, with items that provided the best indicators of the underlying constructs. As a result, the primary goal of the current study was to examine the factorial validity of the instruments to determine whether they provided coherent measures of the latent constructs. Then, given findings of adequate model-data fit, a second goal was to determine how the items in each measurement model rank-ordered in terms of their loadings on the underlying factors. This information would serve to construct the long and short computerised versions of the questionnaires. A final goal of this study was to evaluate the effectiveness of using item-level structural equation modelling as an analytical technique in questionnaire validation.

Given the scale of this project, I will not discuss the findings from every questionnaire. Rather, I will limit the discussion to the Attitudes and Anxiety questionnaires. I will address the following research questions.

1. What is the nature of test-taker attitudes and anxiety as measured by these questionnaires? What is the factorial structure of the Attitudes and Anxiety Questionnaires?
2. How do the items in each questionnaire rank-order from strongest to weakest indicator of the latent factors?
3. How useful is item-level SEM as an analytical procedure for examining the validity of questionnaires?

Study participants

The Attitudes and Anxiety Questionnaires were administered by the EFL Division of UCLES to 207 ESL students studying on summer courses in centres around Britain. In common with the Cambridge *FCE* candidature,

these participants represented a wide variety of native languages, were approximately 15 years of age or older, and had an intermediate or higher level of English language proficiency.

Data collection and preparation

Students were given the Language Attitudes and Anxiety Questionnaire booklets along with an answer sheet. They could take as much time as they needed to complete the questionnaires. Any questions regarding the meaning of the items were answered.

Item-level data were then entered into SPSS version 6.0 for the MAC. Cases with missing data were dropped. The data were then imported into EQS version 5.2 for the MAC for further analyses.

Statistical procedures

Descriptive statistics for each questionnaire item were computed and assumptions regarding normality examined. Items that did not meet these assumptions were considered for removal. Also, items whose means and medians were not within one point of the outer bounds of the scales or whose standard deviation was lower that 1.0 were also considered for removal.

The data were then submitted to a series of reliability analyses. Internal consistency reliability estimates were computed for the individual questionnaire scales. Items with low item-total correlations were dropped or moved to other scales.

Then, each questionnaire was submitted to a series of exploratory factor analyses (EFA) so that patterns in the observed questionnaire data could be examined, and latent factors identified. Items that loaded on more than one variable were flagged for removal, and items that loaded with different sub-scales were considered for change.

Although EFA is a useful statistical procedure for questionnaire validation, it does not have the power of confirmatory factor analysis (CFA). First, EFA procedures assume no a priori patterns in the data, thereby making no a priori constraints on the underlying constructs. In my opinion, this process is out of step with the way the current questionnaires were designed. These instruments were hardly a random assembly of related items. Rather, they were constructed in accordance with a number of principles and studied all along the way. By design, the items were intended to measure one scale and not another. In short, a number of a priori constraints were, in fact, imposed on questionnaires by virtue of the design process. As CFA seeks to determine the extent to which items designed to measure a particular factor actually do so, I felt a confirmatory approach to validation was more appropriate.

Secondly, EFA, as a statistical procedure, is unable to tease apart measurement error from the observed variables, and is unable to determine

whether the errors are correlated. In my experience with questionnaire analysis, the cross-measurement of content as observed through correlated measurement errors can be very common (Purpura 1999). For example, consider the following two items:

(a) I don't like speaking the target language in class (language class anxiety).
(b) I don't like the native speakers of this language (attitudes towards speakers of the target language).

As we can see, the object of measurement in example (a) is 'language class anxiety' and in (b), it is 'attitudes towards speakers of the target language'. Both items, however, are phrased to express a negative preference (I don't like), which might, in fact, introduce construct-irrelevant variance in the measurement. For that reason, CFA is a more appropriate analytic procedure as it allows the researcher to investigate and, if need be, account for correlated errors in the models.

Therefore, following Byrne (1998), each questionnaire was submitted to a series of item-level confirmatory factor analyses (a type of structural equation modelling). Given the fact that respondents were presented with statements rated from 1 to 5 in terms of agreement and were asked to indicate which level most accurately described their opinion, higher scores were assumed to represent a higher level of the underlying factor. These scores, therefore, represented ordinal scale measurement and, according to Jöreskog and Sörbom (1993a, 1993b), should be considered as categorical data. Consequently, these analyses should be based on polychoric correlations, which assume all variables to be measured on an ordinal scale. However, this procedure requires exceedingly large sample sizes, as indicated by the EQS program. For example in analysing the anxiety scale, I was warned that '78,125 cases were needed, but only 207 cases were in the data file'. Because several researchers, including Atkinson (1988), Bentler and Chou (1987) and Muthén and Kaplan (1985), indicate that when the number of categories is large (i.e. four or more), the effect of not treating these data as ordinal may be negligible, I decided to perform two sets of analyses to compare the results. In one set, items were treated as continuous, and analyses were based on a covariance matrix. In the other, items were specified as categorical and analyses were based on a polychoric correlation matrix. The results showed that in the vast majority of cases, the loadings on the underlying factor rank-ordered in the exact same way. In all cases, the model-data fit improved when the items were specified as categorical; however, as the model-data fit statistics for all models were acceptable, this was not an issue for statistical viability. Given the similarities in the results, I will present the results of the data treated as continuous variables.

Model-data fit in this study was evaluated by means of several statistics: the chi-square/degrees of freedom ratio (2 to 1), the comparative fit index (.90 and above), and the root means square error of approximation (RMSEA) (<.05).

Findings

Descriptive statistics

Table 2 presents the summary descriptive statistics for the items in the current study. Most items were within the accepted limits in terms of central tendency, variation and normality. During the course of the analyses, some of the items seen below were dropped and some scales (as indicated below) were merged. Internal consistency reliability estimates for each scale were all in the acceptable range. They are also presented in Table 2.

Table 2 Descriptive statistics
Attitudes towards English Speakers (AES) (N=207)

NAME	MEAN	Std. Dev.	Median	SKEWNESS	KURTOSIS
AES4	2.048	0.729	2.000	0.001	1.632
AES7R	1.749	1.267	2.000	0.223	-1.001
AES12	2.150	0.771	2.000	0.186	1.263
AES16	1.937	0.757	2.000	-0.233	1.375
AES21R	2.604	0.846	2.000	0.082	-0.197
AES23	2.502	0.806	2.000	0.301	-0.180
AES26	2.227	0.825	2.000	-0.183	1.028
AES36	2.329	0.736	2.000	-0.017	0.036
AES37	2.300	0.829	2.000	-0.092	0.253
AES40	2.111	0.712	2.000	-0.001	1.686
AES42	1.889	0.888	2.000	-0.116	-0.138

Alpha = .823 R=recoded item

Attitudes towards Learning English (ALE) (N=207)

NAME	MEAN	Std. Dev.	Median	SKEWNESS	KURTOSIS
ALE1	3.415	0.600	3.000	-0.485	0.641
ALE8	3.106	0.817	3.000	-1.168	2.529
ALE17	3.106	0.702	3.000	-0.914	2.713
ALE20	3.430	0.797	4.000	-1.746	3.689
ALE31	3.179	0.725	3.000	-1.057	2.748
ALE38	2.744	1.160	3.000	-0.748	0.275
ALE14R	3.469	0.774	4.000	-1.857	4.322
ALE28R	3.217	0.896	3.000	-1.671	3.518

Alpha = .720 R=recoded item

Interest in Learning English as a Foreign Language (ILF) (N=207)

NAME	MEAN	Std. Dev.	Median	SKEWNESS	KURTOSIS
IFL6	2.618	1.159	3.000	-0.706	-0.405
IFL9	3.377	0.894	4.000	-1.682	2.772
IFL11	3.242	0.800	3.000	-1.095	1.681
IFL18	2.763	1.018	3.000	-0.626	-0.357
IFL22	3.396	0.835	4.000	-1.664	3.264
IFL30	2.957	0.802	3.000	-0.604	0.724
IFL25R	3.551	0.665	4.000	-2.084	7.176
IFL34R	2.130	1.210	2.000	-0.070	-1.113

Alpha = .600 R=recoded item

Revised Attitudes towards Learning English)(ALE) (N=207)
(merged with Interest in Learning English as a Foreign Language following the analyses)

NAME	MEAN	Std. Dev.	Median	SKEWNESS	KURTOSIS
ALE1	3.415	0.600	3.000	-0.485	-0.641
IFL11	3.242	0.800	3.000	-1.095	1.681
ALE17	3.106	0.702	3.000	-0.914	2.713
ALE28R	3.217	0.896	3.000	-1.671	3.518
ALE31	3.179	0.725	3.000	-1.057	2.748
IFL30	2.957	0.802	3.000	-0.604	0.724

Alpha = .760 R=recoded item

Perception of Task Difficulty (PTD) (N=207)

NAME	MEAN	Std. Dev.	Median	SKEWNESS	KURTOSIS
PTD3	2.159	1.056	2.000	0.026	-0.618
PTD5	2.164	1.076	2.000	-0.049	-0.815
PTD13	1.541	1.060	1.000	0.509	-0.255
PTD15	2.261	1.115	2.000	-0.127	-0.797
PTD24	1.179	0.931	1.000	0.585	-0.099
PTD27	1.261	1.000	1.000	0.604	-0.191
PTD32R	2.087	0.925	2.000	0.308	-0.287

Alpha = .836 R=recoded item

Language Anxiety (N=207)

NAME	MEAN	Std. Dev.	Median	SKEWNESS	KURTOSIS
LANX43	1.691	1.075	1.000	0.335	-0.814
LANX45	1.710	1.116	2.000	0.212	-0.890
LANX51	1.329	1.023	1.000	0.431	-0.599
LANX52	1.502	1.110	1.000	0.349	-0.788
LANX60	1.783	1.041	2.000	0.186	-0.736
LANX63	0.865	0.819	1.000	1.272	2.551
LANX66	2.261	1.162	2.000	-0.166	-0.856
LANX67	1.797	1.139	2.000	0.326	-0.905
LANX71R	0.841	0.824	3.000	0.936	0.861

Alpha = .762 R=recoded item

Class Anxiety (N=207)

NAME	MEAN	Std. Dev.	Median	SKEWNESS	KURTOSIS
CANX50	0.966	0.894	1.000	1.136	1.599
CANX54	0.754	0.843	1.000	1.328	2.068
CANX59	1.348	0.911	1.000	0.460	0.038
CANX62	1.343	0.921	1.000	0.471	0.082
CANX68	1.213	0.883	1.000	0.851	0.903
CANX70	1.710	1.006	2.000	0.146	0.770
CANX72	1.580	1.020	2.000	0.419	0.158
CANX58R	2.585	1.015	3.000	-0.640	0.001
CANX64R	0.952	0.869	1.000	1.035	1.381

Alpha = .552 R=recoded item

Language Class Anxiety (LCANX) (N=207)
(merged with Class Anxiety following the analyses)

NAME	MEAN	Std. Dev.	Median	SKEWNESS	KURTOSIS
CANX50	0.966	0.894	1.000	1.136	1.599
LANX51	1.329	1.023	1.000	0.431	-0.599
LANX52	1.502	1.110	1.000	0.349	-0.788
CANX54	0.754	0.843	1.000	1.328	2.068
CANX62	1.343	0.921	1.000	0.471	-0.082
LANX63	0.865	0.819	1.000	1.272	2.551
CANX72	1.580	1.020	2.000	0.419	-0.158

Alpha = .800

Text Anxiety (TAnx) (N=207)

NAME	MEAN	Std. Dev.	Median	SKEWNESS	KURTOSIS
TANX48	1.657	1.030	1.000	0.432	-0.427
TANX49	1.628	1.001	2.000	0.156	-0.837
TANX53	1.517	1.101	1.000	0.364	-0.769
TANX55	2.034	1.146	2.000	-0.125	-0.966
TANX57	1.802	1.012	2.000	0.009	-0.628
TANX69	1.976	1.063	2.000	-0.000	-0.822

Alpha = .852

The Attitudes Questionnaire

The Attitudes Questionnaire was originally represented as a four-factor model of learner attitudes. It contained four correlated factors: Attitudes toward English Speakers, Attitudes toward Learning English, Interest in Learning English as a Foreign Language, and Perception of Task Difficulty. When each measurement model was estimated separately, the models produced good model-data fit with CFIs greater than 0.95. The individual parameters were sufficiently high, statistically significant and substantively viable, thereby providing good representations of the data.

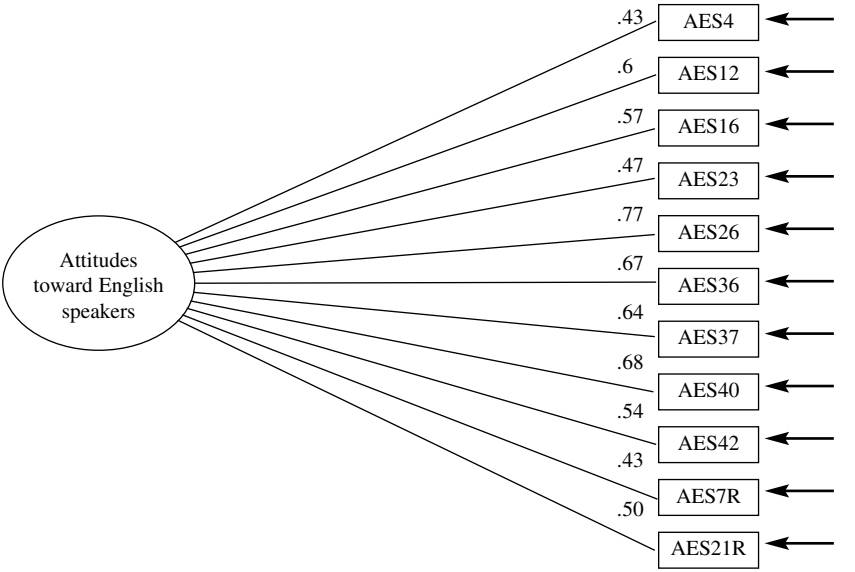
For example, the Attitudes toward Learning English factor in the questionnaire was modelled with eleven items designed as indicators of the latent factor. No correlated errors were hypothesised. The results produced a Chi-square of 64.15 with 44 degrees of freedom, a CFI of .97 and a RMSEA of 0.045. All parameter estimates were statistically significant at the .05 level, and ranged in magnitude from .43 to .77. Thus, this model provided an excellent representation of the questionnaire data, as seen in Figure 1.

Similar results were obtained for the other measurement models in the Attitudes Questionnaire.

However, when the four measurement models were estimated simultaneously, the factor loadings of two scales, ‘Attitudes toward Learning English’ and ‘Interest in Learning English as a Foreign Language’ appeared to be measuring the same underlying construct. Upon reflection, this came as no surprise since, for students studying on summer courses in Britain, interest in learning a foreign language was basically coterminous with interest in learning English. Therefore, these scales were merged as seen in Table 1.

All three measurement models in the Attitudes Questionnaire were then modelled simultaneously. Attitudes were represented by three factors: Attitudes toward English Speakers (AES), measured by eleven items; Attitudes toward Learning English (ALE), measured by six items; and Perception of Task Difficulty (PTD), measured by seven items. No correlated

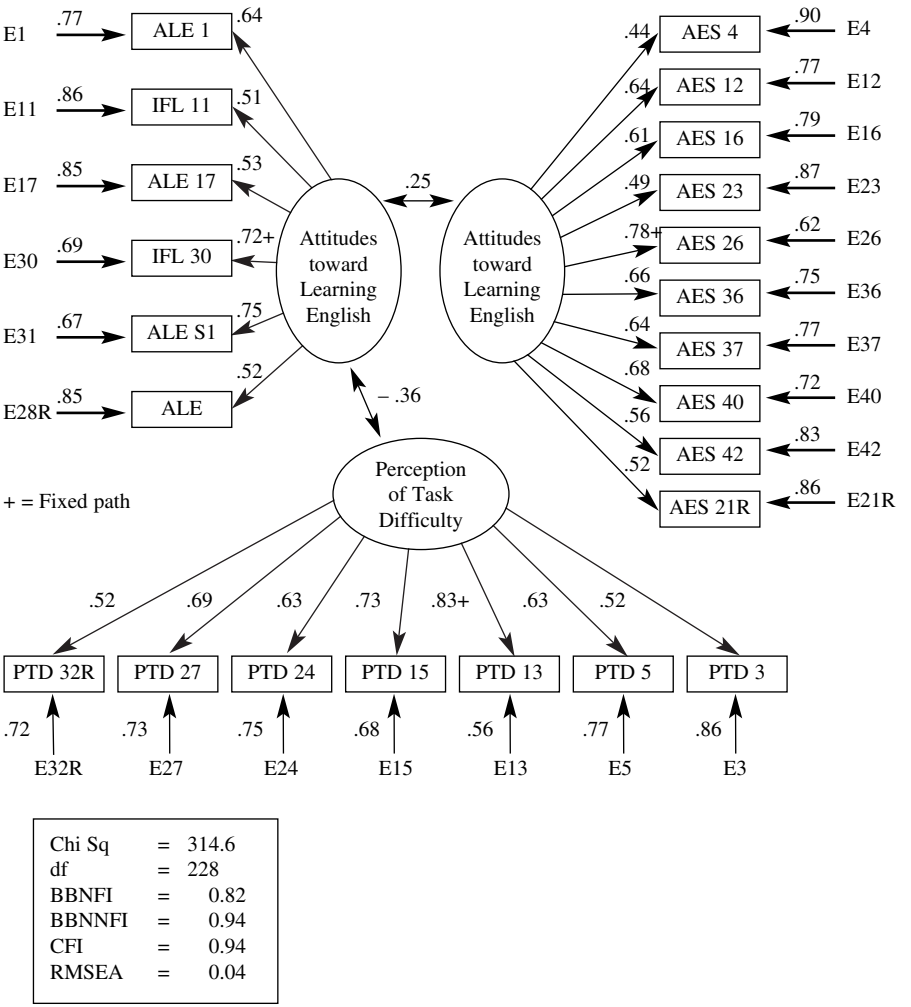
Figure 1 One-Factor Model of 'Attitudes toward English Speakers'



measurement errors were postulated. Based on the work of Gardner (1985a), Attitudes toward Learning English were hypothesised to be correlated with Attitudes toward English Speakers. Similarly, the work of Weiner (1992), Graham (1994) and McCombs (1991) led us to hypothesise a correlation between Attitudes towards Learning English and Perceptions of Task Difficulty. No relationship was hypothesised (and none observed) between Attitudes toward English Speakers and Perception of Task Difficulty. In other words, we did not see the perception of test difficulty as a function of the learners' attitudes toward English speakers. This model of the Attitudes Questionnaire is presented in Figure 2.

The results of this model produced a Chi-square of 314.6 with 228 degrees of freedom, a CFI of .94, and a RMSEA of .04, indicating a good representation of the underlying data. All parameters were statistically significant and substantively viable. This model showed a relatively weak, but statistically significant, relationship (.25) between the students' attitudes towards learning English and their attitudes towards speakers of English. In other words, the students do not seem to be allowing their attitudes towards the speakers of English to affect their attitudes toward learning English. The results also showed a relatively weak negative correlation (-.36) between how difficult students perceived English to be to learn and their affect towards learning English. In short, these results indicated that the more students

Figure 2 A Model of the Attitudes Questionnaire



perceived English to be a hard language to learn, the worse their attitudes tended to become towards learning the language. Given the strength of this association, however, most students appeared not to allow their perception of how hard the language is to affect their attitudes toward learning it.

Based on these results, we were able to use the parameter estimates of each measurement model to rank-order the items from strongest to weakest indicator of the underlying factors. This provided empirical information for

producing both long and short versions of the computerised questionnaires. Table 3 presents the Attitudes toward English Speakers questionnaire along with the factor loadings when the items were treated as continuous and when they were treated as categorical variables. The model-data fit for the former model when the items were treated as continuous was .969 (CFI), and when treated as categorical, it was 1.0. Also, the magnitude of the loadings was lower when treated as continuous, but the rank orderings were the same in both cases.

Table 3 Rank ordering of the AES loadings with items treated as continuous and categorical

CODE	Attitudes Toward English Speakers	(CFI=.969) Loading	(CCFI= 1.0) Loading
AES26	The people who speak this language are friendly.	.772	.827
AES40	The people who speak this language are fun to be with.	.682	.715
AES36	The people who speak this language are welcoming.	.666	.754
AES12	The people who speak this language make good friends.	.642	.672
AES37	The people who speak this language are interesting to talk to.	.639	.669
AES16	The people who speak this language are warm-hearted.	.566	.610
AES42	The people who speak this language are open to people from other cultures.	.541	.594
AES21R	The people who speak this language are boring.	.505	.505
AES23	The more I get to know the people who speak this language, the more I like them.	.468	.468
AES7R	It is difficult to have close friendships with people who speak this language.	.434	.434
AES4	People who speak this language are honest.	.430	.430

The Anxiety Questionnaire

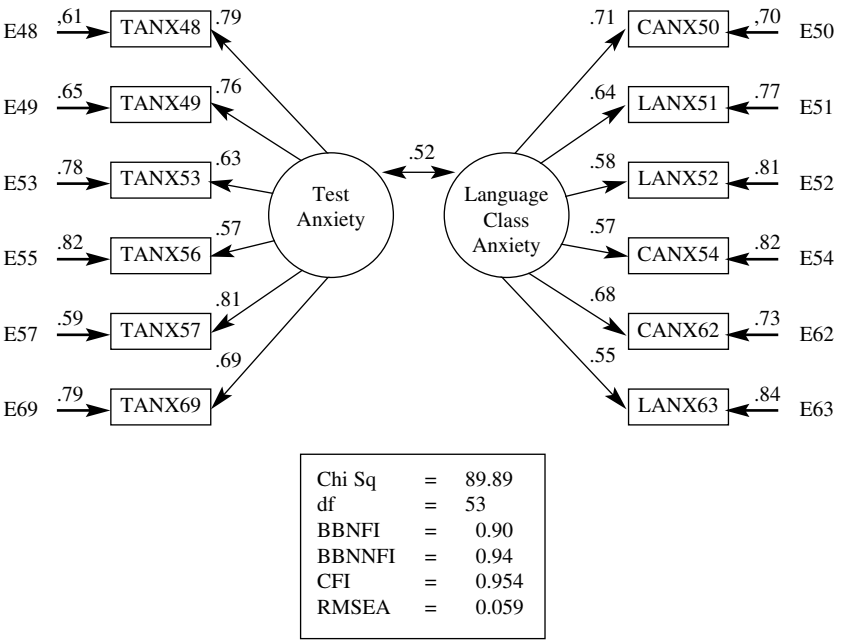
The Anxiety Questionnaire was originally represented as a three-factor model with three inter-correlated factors: language anxiety, class anxiety, and test anxiety. No correlated errors were postulated. When the measurement models were estimated separately, they again produced good model-data fit statistics and the individual parameters were again both substantively and statistically significant.

However, when the three measurement models were estimated simultaneously, the factor loadings of two scales, ‘Language Anxiety’ and ‘Class Anxiety’, again seemed to be measuring one underlying construct. Upon further examination, this may have been because the foreign language was always spoken in class, thereby producing anxiety in some students. For this reason, the two scales were also merged, yielding a two-factor model of anxiety being measured by this questionnaire. These factors included:

language class anxiety (LCanx), measured by six items, and test anxiety (Tanx), measured by six items. These two components of anxiety were hypothesised to be correlated, but again no correlated errors were postulated.

This model produced a Chi-square of 89.89 with 53 degrees of freedom, a CFI of .954, and a RMSEA of .059, indicating that the data provided a good fit for the model. All parameters were statistically significant and substantively viable. The parameter estimates were sufficiently high. This model produced a moderate, statistically significant (at the .05 level) correlation (.52) between language class anxiety and test anxiety, indicating that some test takers who felt anxious in their language classes also tended to feel nervous about language tests. These results are presented in Figure 3.

Figure 3 A Model of the Attitudes Questionnaire



Based on these results, we were again able to use the parameter estimates to rank-order the items in each measurement model from strongest to weakest indicator of the underlying factors. This again allowed us to provide a long and a short version of the questionnaires.

Following these analyses, the following revised taxonomy of socio-psychological factors was produced (see Table 4).

Table 4 Revised taxonomy of the Language Learning Questionnaires

Socio-psychological factors	# of Items
A. Attitudes Questionnaire	**24**
• Attitudes towards English Speakers	11
• Attitudes towards Learning English (merged with Interest in Foreign Languages)	6
• Perception of Task Difficulty	7
B. Anxiety Questionnaire	**13**
• Language Class Anxiety (Language Anxiety was merged with Class Anxiety)	7
• Test Anxiety	6

Conclusion

The purpose of this study was to describe the process used to validate a bank of language learning questionnaires designed to measure selected personal attributes of the Cambridge candidature. Although these procedures were performed on all the questionnaires in the battery, this study reported only on analyses performed on the attitudes and anxiety questionnaires.

The factorial structure of each questionnaire component was modelled separately by means of item-level structural equation modelling. Items that performed poorly were removed. Then, all the components of the questionnaire were modelled simultaneously. With both the attitudes and the anxiety questionnaires, two components appeared to be measuring the same underlying construct. Consequently these components were merged, providing a more parsimonious model of the underlying constructs. Once there was evidence that these models fitted the data well and were substantively viable, the results were used to provide a rank-ordering of the factor loadings associated with each item so that long and short versions of the questionnaires could be produced and subsequently delivered over the computer. In this case, the questionnaires could be customised to provide the strongest indicators of each factor.

These analyses also provided information on the relationship between the underlying factors in the questionnaire. This proved invaluable in fine-tuning the instruments as it allowed us to merge scales when statistically and substantively justifiable. In the attitudes questionnaire, the results showed some correlation between the learners' attitudes towards learning English and their attitudes towards English speakers, while an inverse relationship was observed between the students' perception of how difficult the language was

to learn and their attitudes towards learning English. Then in the Anxiety Questionnaire a moderate relationship was observed between test takers who felt anxious speaking English in class and those who felt anxious taking language tests.

In sum, these results supplied invaluable information on the underlying structure of the questionnaires. Also, by modelling the different components of the questionnaires simultaneously, they provided a better understanding of how the respective components of the questionnaire interacted with one another, providing substantive insights regarding the test takers' personal attributes. In this respect, item-level SEM proved invaluable as an analytical tool for questionnaire validation.

Acknowledgments

Earlier versions of this paper were presented at the 2001 Language Testing Research Colloquium in Saint Louis and at the 2001 ALTE Conference in Barcelona. I would like to thank Nick Saville from UCLES for discussing at these venues how the Language Learning Questionnaires fit into UCLES' Validation plans and how they have been computerised. I am also very grateful to Mike Milanovic and Nick Saville at UCLES for their continued support and encouragement over the year in pursuing this project. Finally, I would like to thank Lyle Bachman and Sara Cushing Weigle for their expertise and inspiration in developing the original version of these questionnaires.

References

Alderman, D. and P. W. Holland. 1981. Item performance across native language groups on the Test of English as a Foreign Language. Princeton: Educational Testing Service.

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

Allan, A. 1995. Begging the questionnaire: Instrument effect on readers' responses to a self-report checklist. *Language Testing* 12: 2, 133–156.

Anderson, N. J., L. Bachman, K. Perkins and A. Cohen. 1991. An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing* 8: 41–66.

Atkinson, L. 1988. The measurement-statistics controversy: Factor analysis and subinterval data. *Bulletin of the Psychonomic Society* 26: 4, 361–364.

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

Bachman, L. F., S. Cushing and J. E. Purpura. 1993. *Development of a questionnaire item bank to explore test-taker characteristics*. Interim Report submitted to UCLES.

Bensoussan, M. and M. Zeidner. 1989. Anxiety and achievement in a multicultural situation: the oral testing of advanced English reading comprehension. *Assessment and Evaluation in Higher Education*, 14: 1, 40–54.

Bentler, P. M. and C. P. Chou. 1987. Practical issues in structural modeling. *Sociological Methods & Research* 16: 78–117.

Brecker, S. J. 1984. Empirical validation of affect, behavior and cognition as distinct components of attitude. *Journal of Personality and Social Psychology* 47: 1191–1205.

Brière, E. 1968. Testing ESL among Navajo children. In J. A. Upshur and J. Fata (eds.), *Problems in Foreign Language Testing. Language Learning, 3*. Ann Arbor, MI: Research Club in Language Learning.

Brown, A. and N. Iwashita. 1996. The role of language background in the validation of a computer-adaptive test. *Melbourne Papers in Language Testing* 5: 1.

Brown, J. D. 1999. The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing* 16: 2, 217–38.

Brown, J. D., G. Robson and P. Rosenkjar. 1996. Personality, motivation, anxiety, strategies, and language proficiency of Japanese students. *University of Hawaii Working Papers in ESL* 15: 1, 33–72.

Byrne, B. M. 1998. *Structural equation modeling with LISREL, PRELIS AND SIMPLIS*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Chapelle, C. 1988. Field independence: A source of language variation? *Language Testing* 7: 121–146.

Chen, C. and Henning, G. 1985. Linguistic and cultural bias in language proficiency tests. *Language Testing* 2: 2, 155–63.

Clapham, C. 1993. Can ESP testing be justified? In D. Douglas and C. Chapelle (eds.), *A new decade of language testing research* (pp. 257–271). Alexandria, VA: TESOL.

Clapham, C. 1996. *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

Clément, R. and B. G. Kruidenier. 1985. Aptitude, attitude and motivation in second language proficiency: A test of Clément's model. *Journal of Language and Social Psychology* 4: 21–37.

Dörnyei, Z. and R. Schmidt (eds.). 2001. *Motivation and second language acquisition*. Manoa, HI: Second Language Teaching and Curriculum Center, University of Hawaii.

Dörnyei, Z. 1990. Conceptualizing motivation in foreign-language learning. *Language Learning* 40: 45–78.

Dörnyei, Z. 2001. *Teaching and researching motivation*. Harlow, England: Longman.

Eiser, J. R. 1986. *Social psychology: Attitudes, cognition and social behaviour*. Cambridge: Cambridge University Press.

Elder, K. 1995. The effect of language background on 'foreign' language test performance. *Language Testing Update* 17: 36–38.

Farhady, H. 1979. The disjunctive fallacy between discrete-point and integrative tests. *TESOL Quarterly* 13: 347–358.

Farhady, H. 1982. Measures of language proficiency from a learner's perspective. *TESOL Quarterly* 16: 1, 43–59.

Fox, J., T. Pychyl and B. Zumbo. 1997. An Investigation of background knowledge in the assessment of language proficiency. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds.), *Current Developments and Alternatives in Language Assessment*, Universities of Tampere and Jyväskylä, Tampere.

Gagné, E. D., C. W. Yekovich and F. R. Yekovich. 1993. *The cognitive psychology of school learning*. Glenview, IL: Scott Foresman. New York: Harper Collins College Publisher.

Gardner, R. C. 1979. Social psychological aspects of second language acquisition. In H. Giles and R. St. Clair (eds.), *Language and Social Psychology* (pp. 193–220). Cambridge, MA: Basil Blackwell, Inc.

Gardner, R. C. 1985a. *Social Psychology and Language Learning: The Role of Attitudes and Motivation*. London, Ontario: Edward Arnold.

Gardner, R. C. 1985b. *The Attitude/Motivation Test Battery: Technical Report*. London, Ontario, Canada: Department of Psychology, The University of Western Ontario.

Gardner, R. C. 1988. The socio-educational model of second-language learning: Assumptions, findings and issues. *Language Learning* 38: 101–126.

Gardner R. C and W. E. Lambert. 1972. Attitudes and motivation in Second Language Learning. Rowley, MA: Newbury House.

Gardner, R. C. and P. C. Smythe. 1975. Motivation and second-language acquisition. *The Canadian Modern Language Review* 31: 3, 218–230.

Gardner, R. C. and P. C. Smythe. 1981. On the development of the Attitudes/Motivation Test Battery. *The Canadian Modern Language Review* 37: 3, 510–525.

Ginther A. and L. Grant. 1997. The influences of proficiency, language background and topic on the production of grammatical form and error in the Test of Written English. In A. Huhta, V. Kohonen, L. Kurki-Suonio, S. Luoma (eds.), *Current Developments and Alternatives in Language Assessment*. Tampere, Finland: Universities of Tempere and Jyväskylä.

Graham, S. 1994. Classroom motivation from an attributional perspective. In H. F. O'Neil Jr and M. Drillings (eds.), *Motivation: Theory and research* (31–48). Hillsdale, NJ Lawrence Erlbaum Associates.

Hansen, J. and C. Stansfield. 1984. Field dependence-independence and language testing: Evidence from six Pacific-Island cultures. *TESOL Quarterly* 18: 311–324.

Hovland C. I. and M. J. Rosenberg (eds.) 1960. *Attitudes, organization and change: An analysis of consistency among attitude components*. New Haven, CT: Yale University Press.

Jensen, C. and C. Hansen. 1995. The effect of prior knowledge on EAP listening-test performance. *Language Testing* 12: 99–119.

Jöreskog, K. G., and D. Sörbom. 1993a. *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Chicago, IL: Scientific Software International.

Jöreskog, K. G., and D. Sörbom. 1993b. *LISREL 8: User's Reference Guide*. Chicago, IL: Scientific Software International. Lawrence Erlbaum Associates.

Kunnan, A. 1990. DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly* 24: 4, 741–746.

Kunnan, A. 1995. *Test Taker Characteristics and Test Performance: A Structural Equation Modeling Approach*. Cambridge: Cambridge University Press.

McCombs, B. L. 1991. Motivational skills training: Combining metacognitive, cognitive and affective learning strategies. In C.E. Weinstein, E. T. Goetz and P. A. Alexander (eds.), *Learning and Study Strategies*. New York: Academic Press, pp. 151–69.

Muthén, B. and D. Kaplan. 1985. A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology* 38: 171–189.

Oltman, P., J. Stricker and T. Barrows. 1988. *Native language, English proficiency, and the structure of the TOEFL*. TOEFL Research Report 27. Princeton, NJ: Educational Testing service.

Oxford, R., M. Nyikos and D. Crookall. 1987. Learning strategies of university foreign language students: A large-scale, factor analytic study. Paper presented at the annual TESOL Convention, Miami, FL.

Purpura, J. E. 1997. An analysis of the relationships between test takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning* 47: 2, 289–325.

Purpura, J. E. 1998. The development and construct validation of an instrument designed to investigate selected cognitive background characteristics of test takers. In. A. Kunnan (ed.), *Validation in Language Assessment* (pp. 111–139). Mahwah, NJ: Lawrence Erlbaum Associates.

Purpura, J. E. 1999. *Learner Strategy Use and Performance on Language Tests: A Structural Equation Modeling Approach*. Cambridge: Cambridge University Press.

Ryan, K. E. and L. F. Bachman. 1990. Differential item functioning on two tests of EFL proficiency. *Language Testing* 9: 1, 12–29.

Skehan, P., 1989. *Individual Differences in Second Language Learning*. London: Edward Arnold.

Skehan, P., 1998. *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Spurling, S. and D. Illyin. 1985. The impact of learner variables on language test performance. *TESOL Quarterly* 19: 283–301.

Stansfield, C. and J. Hansen. 1983. Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly* 17: 29–38.

Sunderland, J. 1995. Gender and language testing. *Language Testing Update* 17: 24–35.

Swinton, S. S. and D. E. Powers. 1980. *Factor analysis of the TOEFL for several language groups*. TOEFL Research Report 6. Princeton, NJ: Educational Testing Service.

Tedik, D. 1990. ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes* 9: 123–144.

Vogely, A. 1995. Perceived strategy use during performance on three authentic listening comprehension tasks. *The Modern Language Journal* 79: 1, 41–56.

Weiner, B. 1979. A theory of motivation for some classroom experiences. *Journal of Educational Psychology* 71: 1, 3–250.

Weiner, B. 1992. *Human Motivation: Metaphors, Theories and Research*. Newbury Park, CA: Sage Press.

Zeidner, M. 1986. Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing* 3: 80–89.

Zeidner, M. 1987. A comparison of ethnic, sex, and age biases in the predictive validity of English language aptitude tests Some Israeli data. *Language Testing* 4: 55–71.

Zeidner, M. and M. Bensoussan. 1988. College students' attitudes towards written versus oral tests of EFL. *Language Testing* 5: 100–114.

6 Legibility and the rating of second-language writing

Annie Brown
University of Melbourne

Introduction

Just as the advent of new technologies, particularly the computer, has had a major impact on the delivery of language programs, with an upsurge in distance-learning programs and independent-learning (CD Rom-based) programs (e.g. Commonwealth of Australia 1999 and 2001), so too is its impact beginning to be seen in increasing use of computers for the delivery of *tests*. Although computers were first used in testing because they allowed for the application of IRT in computer-adaptive tests (e.g. Weiss 1990; Chalhoub-Deville and Deville 1999), more recently they have been used for the delivery of non-adaptive tests also. In the European context in particular, one major European project, DIALANG (e.g. Alderson 2001), aims to deliver a battery of tests via computer for a number of European languages.

Given the widespread interest in computer-based or web-delivered testing (see Roever 2001), it is particularly important to investigate the impact of the technology on test performance. In the context of a move to computer-delivery of TOEFL, Kirsch *et al.* (1998), for example, investigated the effect of familiarity with computers on test takers' performances. This paper is concerned with another aspect of construct-irrelevant variance, namely the relationship of scores awarded on second-language writing tests to essays that have been handwritten vis-à-vis those that have been word-processed.

Handwriting and neatness of presentation has long been seen as a contaminating factor in the assessment of writing ability, and the impact of handwriting on overall judgements of writing quality has been the focus of a number of studies in the area of first-language writing assessment. Some of these studies involved correlations of teacher-assigned ratings of writing quality with independent judgements of handwriting (e.g. Stewart and Grobe 1979; Chou, Kirkland and Smith 1982), whereas others involved experimental designs where the same essays are presented to raters in different presentation formats involving good handwriting, poor handwriting and, in some cases, typed scripts (Chase 1968; Marshall and Powers 1969; Briggs 1970; Sloan and

McGinnis 1978; Bull and Stevens 1979; McGuire 1996). The findings indicate in general that the quality of handwriting does have an impact on scores, and that increased legibility results in higher ratings; in all the studies except that by McGuire (1995), the essays with better handwriting or the typed scripts received higher scores.

Given the great interest over the years in handwriting and its impact on assessments of writing proficiency within the field of *first*-language literacy, it is surprising that there are hardly any studies of the effect of handwriting in the assessment of *second*-language writing. One study involving essays written by non-native speakers (Robinson 1985) produced similar findings to the majority of the first-language writing studies; essays written by students whose L1 did not use the Roman alphabet tended to receive lower scores than essays written by 'expert' writers.

The lack of research into the impact of handwriting on assessments of L2 writing proficiency is all the more surprising in a field where reliability and validity issues are generally well understood, and where much attention is paid in the research literature to identifying and examining the impact of construct-irrelevant variance on test scores. One could argue that it is particularly important in formal L2 writing test contexts to examine and evaluate the impact of extraneous variables such as handwriting and presentation, because it is often on the basis of such tests that decisions regarding candidates' future life or study opportunities are made. Moreover, it is particularly in writing contexts such as these, where writers typically have to write under considerable time pressure, that it may be most difficult for them to control the quality of handwriting and general neatness of layout. It is rare, for example, in formal tests that writers have time to transcribe a draft of the essay into a more legible and well presented script. Also, as Charney (1984) points out, in a test context the constraints imposed on the *rater* may result in handwriting playing a larger part in the assessment than it should. He argues that the assessment constraints – limited time and multiple assessment focuses – mean that raters have to read essays rapidly and this may force them to 'depend on those characteristics [such as handwriting] in the essays which are easy to pick out but which are irrelevant to "true writing ability"'.

It is, perhaps, natural to assume that the same situation would hold for assessments of L2 writing as for L1 writing, that is, that poor handwriting would have a negative impact upon scores. Such an expectation seems logical – a paper that looks good and is easy to read is likely to create a better impression on a rater than one which is messy or difficult to read. Chou *et al.* (1982), for example, point out that crossings-out and re-sequencing of pieces of text may be interpreted as being indicative of a student who is unprepared for writing and unsure of how to sequence his or her ideas; they seem to contend that it may not simply be that poor writing is difficult to process (and therefore assess) but also that raters may make negative inferences about the

character or personality of the writer on the basis of their script untidiness.

There is no obvious reason to suppose that the same features of writing would have a different impact in a L2 writing context. It may even be that students with poor handwriting are even more disadvantaged in L2 contexts because of the centrality of 'fluency' as an aspect of communicative quality. It is difficult to read fluently when illegible handwriting and poor presentation hinder access to the text. Huot (1993) argues that under test marking conditions where rapid reading is required, poor handwriting is likely to impede fluent reading. On the basis of verbal protocol studies of the rating of L2 writing, in fact, it appears that raters do react in much the same way as they do when rating L1 writing; in both contexts it has been found that raters comment frequently and negatively on legibility (see, for example, Huot 1988, 1993; Cumming 1990, 1998; Vaughan 1991; Weigle 1994; Wolfe and Feltovich 1994). Milanovic, Saville and Shen (1996) found that two-thirds of raters involved in their study regarded legibility as having an effect on their scoring.

One could of course argue that, in these days of increased computerisation of tests and test centres, the issue of handwriting in second-language tests is likely to become redundant. However, until then, as long as the two modes are provided as alternatives, the issue of equity arises. The current study arises in the context of a move to administer IELTS in both computerised and pen-and-paper formats, and examines whether the supposed effects do, in fact, occur in second-language contexts; whether typed essays or neatly presented essays receive higher ratings; or whether, given the different rating focus (less on ideas and content and more on the mechanics of writing) or the different test purpose, a different dynamic obtains.

Methodology

The salience of legibility as a factor in raters' judgements was examined within a controlled experimental study in which a comparison was made of scores awarded to scripts which differed only in relation to the variable 'handwriting'. The essay data was gathered using the IELTS Task Two essay.

On the basis of previous studies in L1 contexts (see above), it was hypothesised that scores awarded to the handwritten and typed versions of the essays would be significantly different, with higher scores being awarded to the typed versions. In addition it was hypothesised that the score differences would be greater for those scripts where the handwritten version had particularly poor legibility.

Forty IELTS scripts were selected at random from administrations held at one test centre within a one-year period. The scripts were selected from five different administrations and involved five different sets of essay prompts. Each of the Task Two essays was retyped. Original features such as

punctuation, spelling errors and paragraph layout were retained, but aspects of text editing that would be avoided in a word-processed essay, such as crossings-out, insertions and re-orderings of pieces of text, were tidied up.

Next, in order to produce stable and comparable ratings for the two script types, that is, *handwritten* and *typed* (henceforth H and T), each essay was rated six times. In order to ensure that ratings awarded to an essay in one script type did not affect scores awarded to the same essay in the other format, raters did not mark both versions of the same essay. Rather, each of twelve accredited IELTS raters involved in the study rated half of the typed scripts and half of the handwritten scripts, each being from a different candidate.

Although in operational testing it is left to the discretion of raters as to whether they rate the essays globally or analytically, for the purposes of this study, in order to investigate whether poor legibility had most impact on one particular assessment category, the raters were instructed to assess all the essays analytically. Thus, ratings were awarded to each script for each of the three Task Two analytic categories: *Arguments, Ideas and Evidence, Communicative Quality*, and *Vocabulary and Sentence Structure*. A final overall band score was calculated in the normal way, by an averaging and rounding of the three analytic scores. Raters also took the length of each essay into account in the usual way.

In addition to the IELTS ratings, judgements were made of the legibility of each handwritten script. A six-point scale was developed specifically for the purposes of this study. Drawing on discussions of legibility in verbal report studies such as those discussed above, legibility was defined as a broad concept which included letter and word formation, general layout (spacing, paragraphing and lineation), and editing and self-correction. The four judges (all teachers of writing in first- or second-language contexts) were given written instructions to accompany the scale.

Results

Table 1 shows the mean scores for both the analytic and overall score categories for each version (H and T) of each essay. It shows that both the analytic and the overall scores were on average marginally higher for the handwritten scripts than for the typed scripts. The handwritten scripts achieved a mean rating of 5.30 as opposed to 5.04 for typed scripts for Arguments, Ideas and Evidence (AIE), 5.60 as opposed to 5.34 for Communicative Quality (CQ), 5.51 as opposed to 5.18 for Vocabulary and Sentence Structure (VSS), and an Overall Band Score (OBS), averaged across the three categories, of 5.48 as opposed to 5.17. The spread of scores was similar for both types of script. Although one might expect the score difference to be least marked for Arguments, Ideas and Evidence, as this category is the least concerned with presentation issues, and most marked for

Communicative Quality, as handwriting which is difficult to read will inevitably make the essay less immediately 'communicative', the score difference appears to impact relatively evenly across all assessment categories. In contrast to what was expected, scores were marginally higher for the handwritten scripts than for the typed scripts in all categories.

Table 1 Score analysis of handwritten and typed scripts

Rating Category	Handwritten Mean	SD	Typed Mean	SD	Difference (H-T)	Z	Sig.
AIE	5.30	*0.81*	5.04	0.82	.26	-2.570	.01
CQ	5.60	*0.85*	5.34	0.82	.26	-3.530	.00
VSS	5.51	*0.80*	5.18	0.79	.33	-4.230	.00
OBS	5.48	*0.77*	5.17	0.70	.31	-3.723	.00

In order to investigate the significance of the score differences for the two script types for each rating category, a Wilcoxon matched-pairs signed-ranks test was carried out (see Table 2). As can be seen, although the difference in mean scores is relatively small (0.26 for AIE and CQ, 0.33 for VSS, and 0.27 for OBS), it is nonetheless significant for all rating categories.

The second analysis looked more narrowly at the impact of different degrees of legibility on ratings. On the basis of findings within the L1 writing assessment literature, it was considered likely that the score differences across the two script types (H and T) would be insignificant for highly legible scripts but significant for ones that were difficult to decipher. A comparison was made of the score differences for the ten essays judged to have the best legibility and the ten judged to have the worst (see Appendix for examples of handwriting).

Table 2 Score differences according to handwriting quality

Rating Category	Least legible (n=10)	Z	Sig	Most legible (n=10)	Z	*Sig*
AIE	0.50	-1.84	*.07*	0.13	-1.13	.26
CQ	0.50	-2.20	*.03*	0.15	-1.19	.23
VSS	0.62	-2.67	*.01*	0.17	-.94	.35
OBS	0.59	-2.45	*.01*	0.05	-.20	.83

Table 2 shows the average score difference for the two script types for each set of ten essays. As expected, the score difference between the H and T versions for the candidates with the best handwriting was found to be

relatively small (ranging from 0.05 to 0.17 of a band), whereas for those with the worst handwriting, it was somewhat larger (ranging from 0.5 to 0.62, i.e. at least half a band).

A Wilcoxon matched-pairs signed-ranks test was carried out in order to determine the significance of the score differences between the two script types for each group. For the scripts with the best handwriting, none of the differences were significant. For those with the worst handwriting, AIE was not significant but the other three categories were: CQ at the .05 level, and VSS and OBS at the .01 level.

Discussion

In summary, the analysis found that, as hypothesised, there was a small but significant difference in the scores awarded to typed and handwritten versions of the same essay. Also as expected, the score difference between handwritten and typed essays was greater for essays with poor legibility than for those with good legibility, being on average less than 0.1 of a band for the well written essays but slightly over half a band for the poorly written ones. However, contrary to expectations, it was the handwritten scripts that scored more highly, and the handwritten scripts with poor legibility that had the greatest score difference between versions. In effect, this means that, rather than being *disadvantaged* by bad handwriting and poor presentation, test candidates are *advantaged*.

It is interesting to reflect more closely on why the findings here differ from those found in most studies of first-language writing. As noted earlier, a major difference in the rating of L1 and L2 writing is that in L2 assessments there is a stronger emphasis on mechanics or 'linguistic' features (syntax, grammar and vocabulary) (Cumming 1998). It may be, then, that poor legibility has the effect of masking or otherwise distracting from these sorts of errors. In formal L2 proficiency tests, raters usually have limited time to mark each essay. They also have multiple assessment focuses which demand either multiple readings or a single reading with attention being paid simultaneously to different features. Given this, it may be that the extra effort required to decipher illegible script distracts raters from a greater focus on grammar and accuracy, so that errors are not noticed or candidates are given the 'benefit of the doubt' when raters have to decide between two scores. The corollary of this, of course, is that errors stand out more or are more salient when the essay is typed or the handwriting is clear.

It may also be, of course, that presentation problems inhibit fluent reading to the extent that the quality not only of the grammar, but also of the ideas and their organisation (the coherence of the script), is hard to judge. One rater, for example, commented that she found she had to read essays with poor legibility more carefully in order to ensure that she was being fair to the candidate.

Perhaps when raters do not have the time to read very closely (in operational testing sessions), and the handwriting makes it difficult to read the essay, it may be that they (consciously or subconsciously) 'compensate' and award the higher of two ratings where they are in doubt, in order to avoid discriminating against the candidate. That raters compensate to avoid judging test candidates unfairly appears to be a common finding – they have been found to compensate, for example, for the difficulty of the specific writing task encountered and for perceived inadequacies in the interviewer in tests of speaking.

A third consideration, and one that arose from a later review of the texts judged as the most legible, concerned the 'sophistication' of the handwriting. In the context of this study many of the texts had been produced by learners from non-Roman script backgrounds, and the handwriting in the 'legible' texts, although neat, was generally simple printing of the type produced by less mature L1 writers; indeed it was similar to children's writing (see Appendix). Although there is no evidence of this in what the raters say, it may be that they *unconsciously* make interpretations about the sophistication or maturity of the writer based on their handwriting, which, in the context of a test designed as a university screening test, might affect the scores awarded. This question would require further investigation, as the tests in this study were rated only for neatness, not sophistication.

What this study indicates, other than that poor handwriting does not necessarily disadvantage learners of English in a test context, is that alternative presentation modes are not necessarily equivalent. Given moves in a number of large-scale international tests to ensure that test administrations can be carried out as widely as possible, alternative delivery or completion modes are often considered (see, for example, O'Loughlin 1996). This study shows that, before the operational introduction of alternative modes of testing, research needs to be undertaken into the score implications of such a move in order to determine whether the assumed equivalence actually holds.

Finally, this study indicates a need for further research, not only in terms of replicating the current study (to see whether these findings apply in other contexts) but also in terms of other variables which might arise when a test is administered in two modes. This study, like most of the earlier experimental studies, is concerned with 'directly equivalent' scripts; in investigating the question of legibility it deals only with the same script in different formats; it does not deal with the larger question of differences in composing in the two modes, yet the availability of spell- and grammar-checkers (along with any other sophisticated aids to composing that word-processing software may provide) imposes additional variables which are not within the scope of the present study. It would also be of interest, then, to compare the scores awarded to essays produced in the two formats, pen-and-paper and word-processed, by the same candidates.

References

Alderson, J. C. 2001. Learning-centred assessment using information technology. Symposium conducted at the 23rd Language Testing Research Colloquium, St Louis, MO, March 2001.

Briggs, D. 1970. The influence of handwriting on assessment. *Educational Research* 13: 50–55.

Brown, A. 2000. An investigation of the rating process in the IELTS Speaking Module. In R. Tulloh (ed.), *Research Reports 1999, Vol. 3* (pp. 49–85). Sydney: ELICOS.

Bull, R. and J. Stevens. 1979. The effects of attractiveness of writer and penmanship on essay grades. *Journal of Occupational Psychology* 52: 1, 53–59.

Chalhoub-Deville, M. and C. Deville. 1999. Computer-adaptive testing in second language contexts. *Annual Review of Applied Linguistics* 19: 273–299.

Charney, D. 1984. The validity of using holistic scoring to evaluate writing: a critical overview. *Research in the Teaching of English* 18: 65–81.

Chase, C. 1968. The impact of some obvious variables on essay test scores. *Journal of Educational Measurement* 5: 315–318.

Chou, F. J., S. Kirkland and L. R. Smith. 1982. Variables in college composition (Eric Document Reproduction Service No. 224 017).

Commonwealth of Australia. 1999. *Bridges To China. A web-based intermediate level Chinese course*. Canberra: Commonwealth of Australia.

Commonwealth of Australia. 2001. *Bridges To China CD Rom version. A web-based intermediate level Chinese course*. Canberra: Commonwealth of Australia.

Cumming, A. 1990. Expertise in evaluating second language compositions. *Language Testing* 7: 31–51.

Cumming, A., R. Kantor and D. Powers. 1998. An investigation into raters' decision making, and development of a preliminary analytic framework, for scoring TOEFL essays and TOEFL 2000 prototype writing tasks. Princeton, NJ: Educational Testing Service.

Huot, B. 1988. *The validity of holistic scoring: a comparison of the talk-aloud protocols of expert and novice raters*. Unpublished dissertation, Indiana University of Pennsylvania.

Huot, B. 1993. The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson and B. Huot (eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206–236). Cresskill, NJ: Hampton Press.

Kirsch, I., J. Jamieson, C. Taylor and D. Eignor. 1998. *Computer familiarity among TOEFL examinees: TOEFL Research Report 59*. Princeton, NJ: Educational Testing Service.

McGuire, D. W. 1996. A comparison of scores on the Kansas writing assessment for wordprocessed and hand-written papers of eleventh graders. Dissertation Abstracts International, A: The Humanities and Social Sciences 1996, 56, (9 Mar) 3557-A.

Marshall, J. C. and J. M. Powers. 1969. Writing neatness, composition errors and essay grades. *Journal of Educational Measurement* 6: 2, 97–101.

Milanovic, M., N. Saville and S. Shen. 1996. A study of the decision-making behaviour of composition markers. In M. Milanovic and N. Saville (eds.), *Performance testing, cognition and assessment*: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Studies in Language Testing 3 (pp. 92–114). Cambridge: CUP and UCLES.

O'Loughlin, K. 1996. *The comparability of direct and semi-direct speaking tests: a case study*. Unpublished PhD thesis, University of Melbourne. SILT.

Robinson, T. H. 1985. *Evaluating foreign students' compositions: the effects of rater background and of handwriting, spelling and grammar*. The University of Texas at Austin: Unpublished PhD thesis.

Roever, C. 2001. Web-based language testing. *Language Learning and Technology* 5: 2, 84–94. <hhtp://llt.msu.edu>

Sloan, C. A. and I. McGinnis. 1978. The effect of handwriting on teachers' grading of high school essays (ERIC Document Reproduction Service No. 220 836).

Stewart, M. R. and C. H. Grobe. 1979. Syntactic maturity, mechanics, vocabulary and teacher's quality ratings. *Research in the Teaching of English* 13: 207–215.

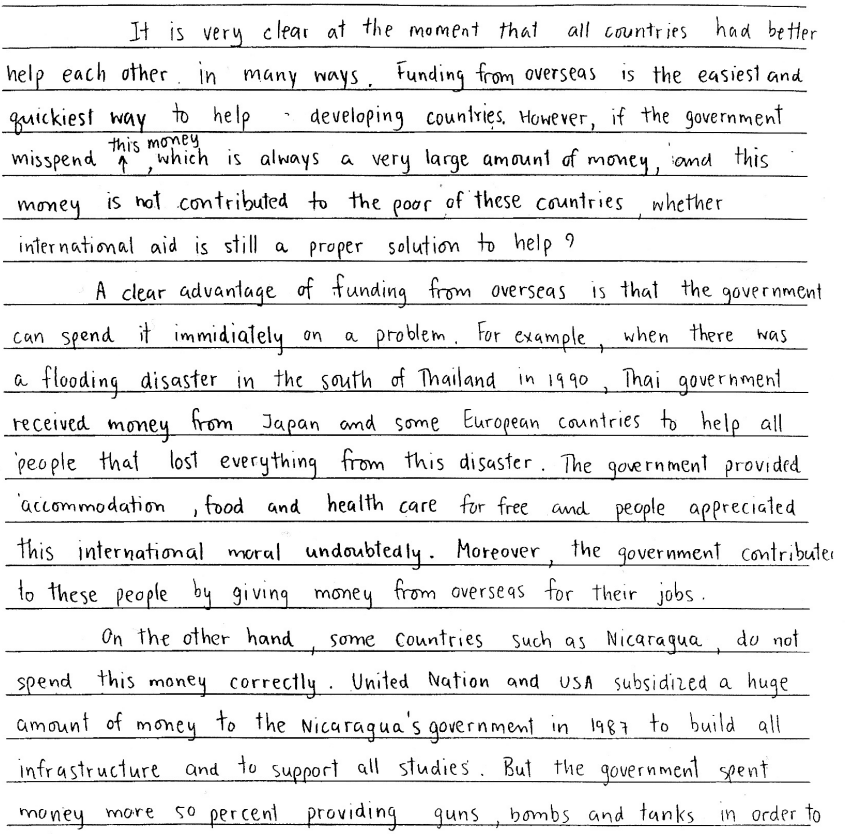
Vaughan, C. 1991. Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (ed.), *Second Language Writing in Academic Contexts* (pp. 111–125). Norwood, NJ: Ablex Publishing Corporation.

Weiss, D. J. 1990. Adaptive testing. In Walberg, H. J. and G. D. Haertel (eds.), *The International Encyclopedia of Educational Evaluation*: 454–458. Oxford: Pergamon Press.

Weigle, S. 1994. Effects of training on raters of ESL compositions. *Language Testing* 11: 2, 197–223.

Wolfe, E., and B. Feltovich. 1994. Learning to rate essays: A study of scorer cognition. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA (ERIC Document Reproduction Service No. ED 368 377).

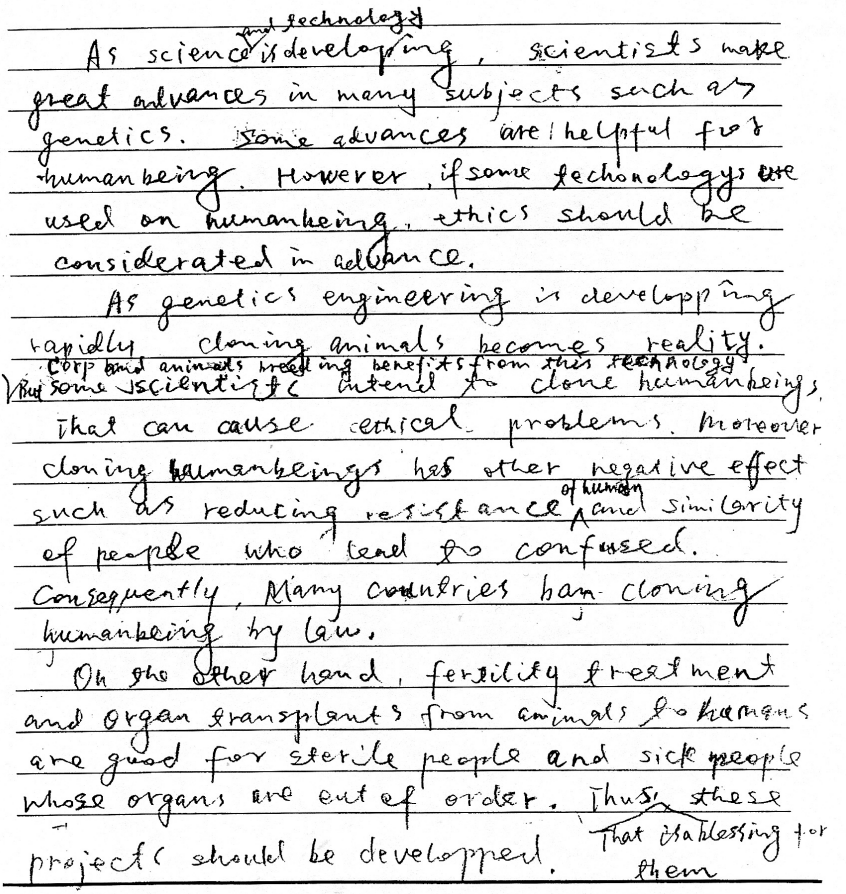
Appendix 1



It is very clear at the moment that all countries had better help each other in many ways. Funding from overseas is the easiest and quickiest way to help developing countries. However, if the government misspend this money, which is always a very large amount of money, and this money is not contributed to the poor of these countries, whether international aid is still a proper solution to help?

A clear advantage of funding from overseas is that the government can spend it immidiately on a problem. For example, when there was a flooding disaster in the south of Thailand in 1990, Thai government received money from Japan and some European countries to help all people that lost everything from this disaster. The government provided accommodation, food and health care for free and people appreciated this international moral undoubtedly. Moreover, the government contributed to these people by giving money from overseas for their jobs.

On the other hand, some countries such as Nicaragua, do not spend this money correctly. United Nation and USA subsidized a huge amount of money to the Nicaragua's government in 1987 to build all infrastructure and to support all studies. But the government spent money more 50 percent providing guns, bombs and tanks in order to


As science and technology is developing, scientists make great advances in many subjects such as genetics. Some advances are helpful for human being. However, if some technologys are used on humankeing, ethics should be considerated in relevance.

As genetics engineering is developping rapidly cloning animals becomes reality. Corp and animals reading benefits from this technology But some scientists intend to clone humankeings, that can cause ethical problems. Moreover cloning humankeings has other negative effect such as reducing resistance of human and similarity of people who tend to confused. Consequently, Many countries ban cloning humankeing by law.

On the other hand, fertility treatment and organ transplants from animals to humans are good for sterile people and sick people whose organs are out of order. Thus, these projects should be developped. That is blessing for them

7 Modelling factors affecting oral language test performance: a large-scale empirical study

Barry O'Sullivan
University of Reading

Background

O'Sullivan (1995, 2000a, 2000b, 2002) reported a series of studies designed to create a body of empirical evidence in support of his model of performance (see Table 1 for an overview of these studies). The model sees performance on oral proficiency tests (OPTs) as being affected by a series of variables associated with the test taker, the test task and the interlocutor. The first three studies, included in Table 1, focused on particular variables, isolated under experimental conditions, and found mixed evidence of significant and systematic effects. An additional study (O'Sullivan 2000b), designed to explore how the variables might interact in an actual test administration, failed to provide evidence of any significant interaction.

The study reported here was designed to further explore the effect of interactions amongst a series of variables on OPT performance (as represented by the scores awarded). Therefore, the hypothesis tested in this study can be stated as

> In a language test involving paired linguistic performance, there will be a significant (α<.05) and systematic interaction between the variables relative age and gender, acquaintanceship, perceived relative personality and perceived relative language ability.

Method

The Test takers

The results for a total of 565 candidates from three major European test centres (Madrid, Toulouse and Rome) on the Cambridge First Certificate in English (FCE) were included in this study. The candidates were representative of the typical FCE population in terms of age, gender, educational background and test experience, based on data provided from the Candidate Information Sheets (CIS) completed by all UCLES Main Suite candidates.

Table 1 A summary of background studies

Variable(s) studied	Participants	Task(s)	Analysis	Results
Age (O'Sullivan 1995)	16 Japanese males 'Young' group (20–24) 'Old' group (36–48)	1. Interview – once by a similarly-aged interviewer, once by an older or younger interviewer 2. Pair-work – once with similarly-aged partner, once with older or younger partner	Multi-faceted Rasch bias interaction analysis	No clear patterns of bias towards or against either group for a particular task type
Gender (O'Sullivan 2000a)	12 Japanese university students (6 men and 6 women, average age approx. 20)	Structured interview format Part 1 short answers Part 2 longer responses Interviewed twice, once by a woman and once by a man	Quantitative data (FSI scale) – two-way repeated measure ANOVA Qualitative data – accuracy & complexity (participants' language); speech characteristics (interviewer language)	Sig. difference found – higher when interviewed by a woman (though the Grammar criterion was the only sig.) Sig. diff. found for accuracy (but not complexity) Sig. diff. found in the language of the interviewers
Acquaintanceship (O'Sullivan 2002)	Phase 1 12 Japanese women (aged 21–22)	Personal info. exchange Narrative based on set of pictures Decision-making task All performed once with friend, once with stranger	Quantitative – Wilcoxon Matched-Pairs Signed-Ranks on FSI scores	Sig. diff. found (higher with friend) – actual difference of almost 10% (discounting two outliers)
	Phase 2 24 Japanese (16 women, 8 men, 20–22) two institutions	Same Tasks as Phase 1	Quantitative – Repeated measures ANOVA for FCE scores Qualitative – accuracy & complexity of participants' language	Sig. diff. found (higher with friend) though sex of stranger not sig Sig. diff. found in the accuracy (but not complexity)
Multi-Variable Study (O'Sullivan 2000b)	304 Turkish students (148 women, 156 men)	Task 1 – Personal information exchange Task 2 – Pair-work (selection of items for holiday – graphic given) Task 3 – Negotiation for additional items	Quantitative – MANOVA, General Linear Model on FCE scores, using responses to questionnaire (perception of partner) Analysed in 2 phases.	Phase 1 – Sig. main effect for Partner Sex & Partner Acquaintanceship (no interaction) higher when partner is male, and a stranger Phase 2 – no sig. diff. observed for Personality

The candidate questionnaire

In addition to the test score and CIS data, a brief questionnaire was also given to all candidates. The items on the questionnaire referred to the gender of the candidate and of their interlocutor, in addition to the candidate's perception of:

- the age of their partner relative to their own age
- the level of proficiency of their partner relative to their own level of proficiency
- the personality of their interlocutor compared to their own (limited here to intraversion/extraversion)
- the degree of acquaintanceship between the candidate and his/her interlocutor.

The examiners

In all, a total of 41 examiners took part in the test administration over the three sites. All examiners were asked to complete an information sheet on a voluntary basis, and all examiners who participated in this administration of the FCE did so. The data collected indicated an even male/female divide, and suggested that the typical examiner was an English teacher (most commonly with a Diploma-level qualification), in the region of 40 years old, and with extensive experience both as a teacher and of the FCE examination.

The design of the study

The design of this phase of the project focused on the probability of variable performance under examination conditions. This meant that test takers were assessed in dyads by a pair of testers, one acting as participant and the other acting as the observer.

Data collection and management

All questionnaire data were input to a data spreadsheet (Microsoft Excel) and merged with an existing data set which included demographic data routinely collected using CIS sheets and test-score data.

Data analysis

A measure of inter-rater reliability was first established by using correlations. Cross-tabulation tables were then used in order to establish that there were adequate individual sub-populations for the main analysis. ANOVA was then carried out using the MINITAB statistical package.

Inter-rater reliability

Analysis of the data using SPSS indicated that the Pearson correlation coefficient was .78. This figure indicates that there is a significant degree of overlap between the two ratings – but also that there are sufficient differences

to support the argument that they are both either measuring a different aspect of oral proficiency or are measuring the same ability, although from different perspectives. Both of these figures are significant at the .01 level, and should be considered satisfactory for this type of 'live' test.

Results

Before performing the main analysis, it was first necessary to explore the data from the Candidate Questionnaires (CQs), in order to establish the final population for the proposed analysis. Table 2 contains the responses to the items on the Candidate Questionnaire (CQ). From this table we can see that the population mix is approximately 55% female and 45% male – this mix is reflected in the 'partner gender' figures. Any difference in number is a reflection of the fact that not all test takers in all centres are included in this population – for instance where they have not fully completed the CIS or CQ.

It is clear from this table that the five-level distinction for items does not yield sufficiently large cell sizes for analyses across all levels. For this reason, it was decided to collapse the extremes for each variable. The results of this procedure are included in the table, in the rows entitled '3-levels'.

Table 2 Cross-tabulation of candidate questionnaire responses

Item	Variable	Responses				
1	Candidate Gender	Women **314**	Men **252**			
2	Partner Gender	Women **320**	Men **246**			
3	Partner Age	Much younger	Younger	Similar	Older	Much older
	5-levels	17	128	290	123	8
	3-levels	**145**		**290**	**131**	
4	Partner Personality	Much less outgoing	Less outgoing	Similar	More outgoing	Much more outgoing
	5-levels	6	54	396	101	9
	3-levels	**60**		**396**	**110**	
5	Partner Lang. Level	Much lower	Lower	Similar	Higher	Much higher
	5-levels	1	45	400	116	4
	3-levels	**46**		**400**	**120**	
6	Acquaintanceship	Stranger	To see	Acquaint.	Friend	Close friend
	5-levels	234	20	138	133	41
	3-levels	**254**		**138**	**174**	

In Item 4, the response is not unexpected, as we would normally expect to find very few test takers (or members of any other group representative of a wider population) who would be classified as being extreme, even when measured by formal estimators of personality profile such as the Meyers-Briggs Indicator or the Eysenck Personality Questionnaire. Thus the fact that only 2.7% of the test takers saw their partners as being extremely different from them in terms of extraversion can be seen as indicating that this item is reflecting test-taker perceptions quite accurately – and suggests that these perceptions are also accurate. With Item 5 we would not expect a great spread of proficiency level; after all, these test takers indicate that they are familiar with the UCLES Main Suite Examinations (to the extent that many of them had previous experience of these or other UCLES tests) and have elected to sit for the FCE. Of course some spread of ability level is bound to be found in a population of this size, and the natural tendency will be for learners to overestimate the ability of their peers – see Horowitz's (1986) critique of the process approach when applied in an academic writing situation, and Berg's (1999) suggestion of the need for training prior to any peer response activity.

ANOVA Results

The decision to perform the initial ANOVA on the Overall Total Score (OVTOT) was taken as this score represents the reported score, and, as such, is the most important as far as test outcome is concerned. Through this ANOVA it should be possible to identify any main effects and/or interactions among the six independent variables – thus identifying variables or combinations of variables which appear systematically to affect performance (as represented by the score achieved).

ANOVA – Overall Total Score

Table 3 Analysis of variance for overall total score

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Age3*Pers3	4	46.387	101.615	25.404	2.92	0.021
candgen*partgen*LLev3	2	26.851	57.403	28.702	3.30	0.038
candgen*Pers3*Acq3	4	69.869	93.675	23.419	2.69	0.031
partgen*Age3*Acq3	4	77.986	94.781	23.695	2.72	0.029
candgen*partgen*Age3* Pers3	4	18.560	25.531	6.383	0.73	0.569
candgen*partgen*LLev3* Acq3	4	22.225	28.296	7.074	0.81	0.517
candgen*partgen*Age3*LLev3	4	36.124	29.260	7.315	0.84	0.500
candgen*partgen*Age3*Acq3	4	64.813	64.813	16.203	1.86	0.116
Error	437	3801.041	3801.041	8.698		
Total	564	5028.796				

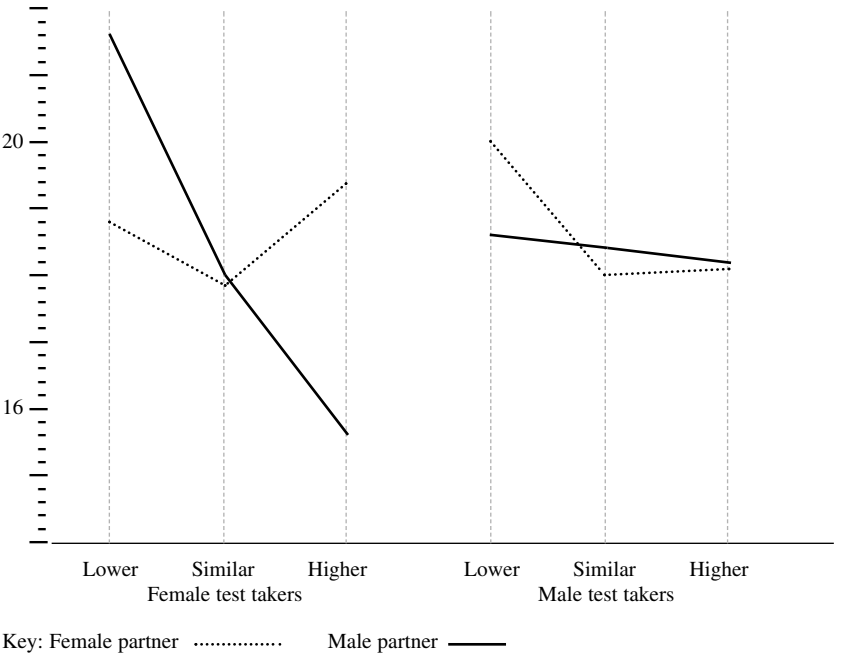
The results of this ANOVA, a summary of which is shown in Table 3, indicates that there are significant three-way interactions on three occasions, and these are:

- Candidate Gender * Partner Gender * Partner Language Level
- Candidate Gender * Partner Personality*Acquaintanceship
- Partner Gender * Partner Age * Acquaintanceship

Candidate Gender * Partner Gender * Partner Language Level

The interaction plot for this effect is shown in Figure 1 below. From this figure it is clear that there are a number of interactions taking place, and that the effect highlighted by the ANOVA appears to be both statistically significant and of practical importance – as the range of mean scores appears very wide.

Figure 1 Interaction plot for overall total score (3-way interaction) – Candidate Gender*Partner Gender*Partner Language Level



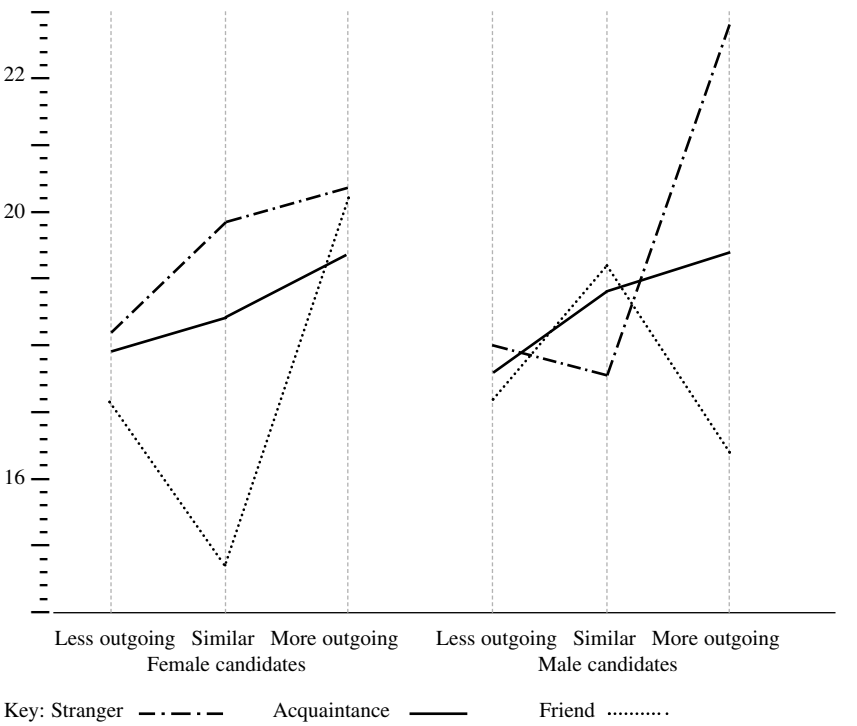
From this interaction plot we can see that there is a clear difference between the male and female test takers. While there is a very slight fall in the mean score of male candidates relative to the perceived language level of their male interlocutors, they appear to be achieving considerably higher scores when paired with a woman whose language level they perceive to be lower than their own (almost 2 points out of a possible 25 – approximately 8% of the

range). On the other hand, the men appear to achieve similar scores with women whom they consider to be at either the same or a higher language level than themselves. Female test takers, on the other hand, seem to display wider, and more significant, differences under the different conditions. Unlike their male counterparts, the difference in mean scores achieved when working with other female test takers at the three different language levels varies by up to 1.5 points (6% of the range). However, it is with male partners that the scores really show a dramatic degree of variation. Here, there appears to be a systematic lowering of the mean score as the perceived language level of the male partner increases. The range of difference is approximately 6 points, or 24% of the possible range.

Candidate Gender * Partner Personality * Acquaintanceship

In the graphic representation of this interaction (see Figure 2 below) we can see that the male and female test takers tend to achieve similar patterns of mean scores when working with an acquaintance – with similar differences in scoring range also (approximately one point or 4% of the overall). However,

Figure 2 Interaction plot for overall total score (3-way interaction) – Partner Gender*Partner Personality*Acquaintanceship



when it comes to the other conditions, it is clear from the graph that there are very different patterns for the male and female candidates.

The data suggest that there is relatively little difference in mean scores when the female test takers are working with a partner whom they perceive as being more outgoing than themselves, irrespective of the degree of acquaintanceship. The same can be said of partners perceived as being less outgoing than themselves – though there is a clear tendency for these test takers to achieve scores that are approximately 2 points (8%) higher when working with partners perceived as being more outgoing.

It is when working with a partner considered to be a friend, similar in personality to themselves, that the mean scores achieved by the female test-takers are the most variable – with an overall range of mean scores of 5.5 points (22% of the overall score). In order to double-check that the data upon which this chart is based were reliable, a review of the original data set was made at this point. Although there was one instance of a very low-scoring test taker among this group, the overall effect does not change dramatically if that score is removed.

In contrast, the male test takers seem to achieve similar scores when working with partners they perceive as less outgoing or similar in personality to themselves, regardless of the degree of acquaintanceship. Here, the greatest variation appears to be when the partner is perceived as being more outgoing – with a systematic increase in score of approximately 6 points (24%) from a low of 16.52 with a friend, to 19.46 with an acquaintance, to a high of 22.72 with a stranger.

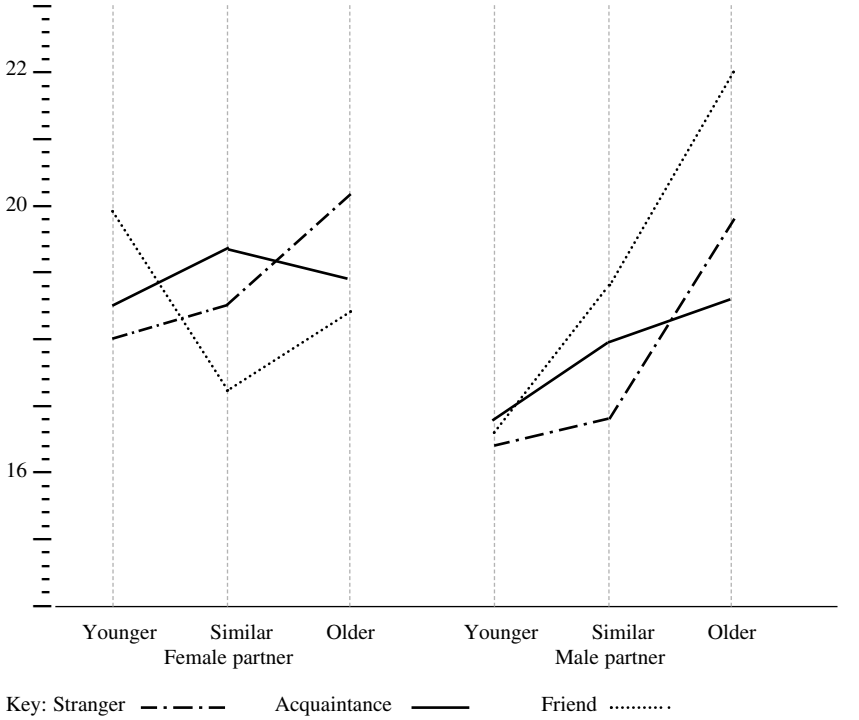
Partner Gender * Partner Age * Acquaintanceship

In the final 3-way interaction there is again a clear difference in mean scores awarded under the different conditions (see Figure 3).

There appears to be a certain systematicity to the scores achieved by the test takers when working with a male partner. While there is little difference when the partner is younger than the test taker (irrespective of the degree of acquaintanceship), there is some considerable difference when the partner is older. Though this is true of all conditions, the difference in mean score between working with a younger friend and working with an older friend is approximately 5 points (20%). We can also say that the clearest distinction between working with a male stranger, acquaintance or friend comes where that person is seen to be older.

With the female partners the picture is far more complicated. While there appears to be little difference in performance with an acquaintance (irrespective of their age), there appears to be a degree of systematicity in the rise in mean score when test takers are paired with a female stranger, resulting in the highest mean scores when they are paired with an older stranger. Looking across the graph we can see that the same pattern of mean scores can

Figure 3 Interaction plot for overall total score (3-way interaction) – Candidate Gender*Partner Age*Acquaintanceship



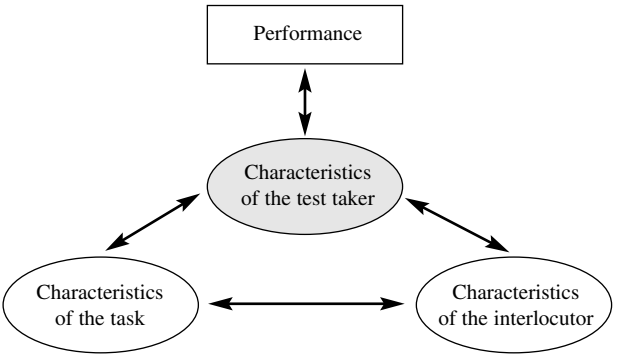
be found with both male and female partners – though the mean scores for the former are approximately almost 2 points higher than those for the latter. The greatest differences between the two sides of the graph are to be found in the interactions with a friend. Whereas the test takers appeared to gain systematically higher scores with older male strangers, when they are paired with a female test taker they appear to achieve the highest mean scores when they consider that person as being younger than themselves.

Summary and initial conclusions

The results of this study suggest that there are particular combinations of variables (which represent one characteristic of the test taker and five characteristics of the interlocutor – seen from the perspective of the test taker) which, when combined in a number of ways, tend to result in particular scoring patterns among the test takers who participated in this study. This implies that there is empirical support for the hypothesised interaction between the variables described in this study. The fact that all six variables

(again in particular interactions) seem to have had some effect on the mean scores achieved by these test takers, confirms that the model of performance proposed by O'Sullivan (2000b), and reproduced here as Figure 4, may be used as a starting point for continued exploration of the notion of performance. When consideration was given to the weighting of test-taker scores used by Cambridge, ANOVA indicated that the same interaction effects were to be found as were found in the analysis using the unweighted scores, though the balance of the contribution of the scores from the different examiners was altered somewhat.

Figure 4 Model of performance (O'Sullivan 2000b)



It must, of course, be noted that the operationalisation of the notion of performance in this study is based on the scores awarded by trained examiners based on their perception of the success or failure of those performances. It is possible that the scores awarded here may have been affected by a number of factors. The factors are: characteristics of the examiners, characteristics of the task, characteristics of the assessment criteria, interactions among these characteristics, and affective reactions of the examiners towards the test takers.

Additional conclusions

- The results suggest a link between performance on a test (as seen through the eyes of the examiners) and the test takers' perceptions of the person they are paired with in that test. The conclusion must be that test takers' reactions to their partner somehow affect their performances on tasks.
- There is a distinction made in the model between the interlocutor and the task. It can be argued that this distinction is not entirely valid, as the interlocutor is an aspect of the test performance conditions. The distinction

is made here in order to reflect the important role played by the interlocutor in any interaction.

- The model presented here can be seen as a move towards a more socio-cognitive view of performance in which the cognitive processing of certain kinds of information is recognised as being socially driven (see Channouf, Py and Somat (1999) for evidence of empirical support for this approach to cognitive processing).

Implications for examinations boards

a) Consideration might be given to allow test takers to choose their own partner, as they are likely to opt for a person with whom they feel they can produce their optimum performance, though there are generalisability issues inherent in allowing for this kind of pair self-selection.
b) In the event of this not being possible, the board should draw up a list of partner qualities in the form of a set of empirically derived guidelines, which will result in optimum performance from both test takers. These guidelines would then be used by examination centres as an aid to the selection of pairings.

Implications for language testing in general

1. A number of researchers distinguish between test features that are irrelevant to the ability that is being measured, and those which are relevant to that ability (Locke 1984; Porter 1991a, 1991b; Porter and Shen 1991; O'Sullivan 1995; O'Sullivan and Porter 1995). If a feature affects test results to a significant degree, but is irrelevant to the ability being measured, it is indeed a source of measurement error which needs to be eliminated. If it is relevant to the ability being measured, however, and occurs in tests because it is an essential and naturally-occurring part of natural language use, and if it affects test results to a significant degree, it is desirable that it should be included in test activities and tasks.

Figure 5 Predictability and relevance of characteristics

Predictable Irrelevant	Predictable Relevant
Unpredictable Irrelevant	Unpredictable Relevant

Some of the characteristics are going to be predictable/unpredictable and/or relevant/irrelevant. While for some variables the position in the matrix will remain unchanged, irrespective of the task, for others the position will be different.

2. McNamara (1997: 458) sees research on the impact of interlocutor behaviour on performance as having implications for 'interlocutor training and interlocutor selection'. There have been a number of studies which have demonstrated that there is variability in the performance of the examiner (for example, Lazaraton 1996a, 1996b; Brown 1998; O'Sullivan 2002). It would appear then, that we have enough evidence to say that in certain situations it is likely that there will be an interlocutor effect on performance. We have yet to discover *why* this is happening and *how* exactly it is manifested in the language of the interaction.
3. Rater training is described by Bachman and Palmer (1996: 221) as being '[O]ne of the most effective ways of dealing with inconsistency'. The problem with the phenomenon reported here is that it is not associated with rater inconsistency, but with the language used by the candidates, and as such cannot be dealt with by rater or examiner training (though it may be possible to raise examiners' awareness of the danger of such affective reactions on the part of candidates and to train examiners to recognise when it is happening and to deal with it by deliberate intervention in the interaction).
4. One final implication for oral language testing concerns the idea of 'bias for best' (Swain 1985: 42–43). These studies appear to cast some doubt on the feasibility of operationalising this notion for this type of test. This is because it would appear that we neither know what best is, nor can we say that 'best' for one test taker will be 'best' for another.

Implications for language teaching, learning and research

The work of a group of researchers in the 1980s and early 1990s was focused on the benefits of using tasks that promote negotiation of meaning and language learning (a point they argued but never demonstrated empirically) among pairs and small groups of language learners. The evidence from this series of research studies suggests that most of these studies would benefit from replication with a more explicit awareness of the characteristics of the participants. This is of particular importance as the studies in question represent a body of research which forms the central theoretical basis of task-based learning.

References

Bachman, L. F. and A. S. Palmer. 1996. *Language Testing in Practice*. Oxford: OUP

Berg, E. C. 1999. The effects of trained peer response on ESL students' revision types and writing quality. *Journal of Second Language Writing*. Vol. 8: 3, 215–241.

Brown, A. 1998. Interviewer style and candidate performance in the IELST oral interview. Paper presented at the Language Testing Research Colloquium. Monterey CA.

Channouf, A., J. Py and A. Somat. 1999. Cognitive processing of causal explanations: a sociocognitive perspective. *European Journal of Social Psychology* 29: 673–690.

Horowitz, D. 1986. Process, not product: Less than meets the eye. *TESOL Quarterly* 20: 141–144.

Lazaraton, A. 1996a. Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing* 13: 2, 151–172.

Lazaraton, A. 1996b. A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE). In M. Milanovic and N. Saville (eds.) *Performance Testing, Cognition and Assessment: selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*. Studies in Language Testing 3, pp. 18–33.

Locke, C. 1984. *The influence of the interviewer on student performance in tests of foreign language oral/aural skills*. Unpublished MA Project. University of Reading.

McNamara, T. F. 1997. 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics* 18: 4, 446–466.

Milanovic, M. and N. Saville. 1996. Introduction. Performance Testing, Cognition and Assessment. Studies in Language Testing 3. Cambridge: UCLES, 1–17.

O'Sullivan, B. 1995. *Oral language testing: Does the age of the interlocutor make a difference*? Unpublished MA Dissertation. University of Reading.

O'Sullivan, B. 2000a. Exploring Gender and Oral Proficiency Interview Performance, *SYSTEM* 28: 3, 373–386

O'Sullivan, B. 2000b. *Towards a model of performance in oral language testing*. Unpublished PhD dissertation, the University of Reading.

O'Sullivan, B. 2002 Learner Acquaintanceship and OPI Pair-Task Performance, *Language Testing*, Volume 19: 3, 277–295.

Porter, D., & O'Sullivan, B. (1999). The effect of audience age on measured written performance. System, 27, 65-77.

Porter, D. 1991a. Affective Factors in Language Testing. In Alderson, J. C. and B. North (Eds.) Language Testing in the 1990s. London: Macmillan (Modern English Publications in association with The British Council), 32-40.

Porter, D. 1991b. Affective Factors in the Assessment of Oral Interaction: Gender and Status. In Sarinee Arnivan (ed) Current Developments in Language Testing. Singapore: SEAMEO Regional Language Centre. Anthology Series 25: 92-102

Porter, D. and Shen Shu Hung. 1991. Gender, Status and Style in the Interview. The Dolphin 21, Aarhus University Press: 117-128.

Swain, M. 1985. Large-scale communicative language testing: A case study. In Y. P. Lee, A. C. Y. Y. Fok, R. Lord and G. Low (eds.) *New Directions in Language Testing*. Oxford: Pergamon. pp. 35–46.

8 Self-assessment in DIALANG An account of test development

Sari Luoma
Centre for Applied Language Studies,
University of Jyväskylä, Finland

This article reports on the development of the self-assessment section in DIALANG. Like any test development process, this started with initial development and proceeded to piloting, which led to revisions, more piloting, and preparations for publishing Version 1 of the system as well as plans for new developments. The article begins with a brief introduction to the DIALANG system and the role of self-assessment in it. The main part of the paper then reports on the development of the self-assessment section, the basis for the development decisions, and the steps foreseen in the near future for further work on self-assessment in DIALANG. The conclusion emphasises the importance of combining qualitative and quantitative investigations in test development.

What is DIALANG?

DIALANG is a European project that is developing a computer-based diagnostic language assessment system for the Internet. The project has been going on since 1997, currently with 19 institutional partners, with financial support from the European Commission Directorate-General for Education and Culture (SOCRATES programme, LINGUA Action D).

The upcoming DIALANG system has unprecedented linguistic coverage: it has 14 European languages both as test languages and as support languages, i.e. languages in which users can see the program instructions. DIALANG is intended to support language learners, and, in addition to test results, it offers users a range of feedback and advice for language learning. While it proposes a certain assessment procedure, the DIALANG system allows users to skip out of sections and to select which assessment components they want to complete in one sitting.

The assessment procedure in Version 1 of the DIALANG system begins

with a test selection screen. Here, the user chooses the test he or she wants to take, including choice of support language, the language to be tested, and the skill to be tested. There are five skill sections available in DIALANG: listening, reading, writing (indirect), vocabulary, and structures. Thus, the user can choose, for instance, a listening test in French with instructions in Greek, or a vocabulary test in German with instructions in English.

The selected test then begins with a placement test and a self-assessment section. The combined results from these are used to select an appropriate test for the user. The self-assessment section contains 18 'I can' statements and, for each of these, users click on either Yes (=I can do this) or No (=I can't do this yet). The development of this section will be discussed in this article. Once both sections are completed, the system chooses a test of approximately appropriate difficulty for the learner, and the test begins.

In DIALANG Version 1, the test that the user gets is fixed and 30 items long. In Version 2, the DIALANG system will be item-level adaptive, which means that the system decides which items to give to the user based on his or her earlier answers. If the users get an item right, they get a more difficult item. If they get it wrong, they get an easier item. When the system is fairly sure that a reliable assessment of the learner's proficiency has been reached, the test ends, probably in less than 30 items. Once the tests are fully adaptive, the placement test at the beginning of the procedure will no longer be required, but the self-assessment section will remain a part of the DIALANG system because of the belief that it is beneficial for language learning.

For the DIALANG user, the test works in much the same way whether it is Version 1 or Version 2. The user gets test items in the selected skill one by one, and when the test is complete, the user gets feedback.

The feedback menu in DIALANG has two main headings, 'Results', and 'Advice'. Under 'Results', the users have access to their test results including comparison between their self-assessment and their test-based proficiency level, their results from the placement test, and an option to go back and review the items that they answered. They can see whether they got each item right or wrong, and what the acceptable answers were. Under 'Advice', the users can read advice about improving their skills, and explanations about self-assessment. This may be particularly interesting for users whose test score did not match their self-assessment. The users may read as much or as little feedback as they wish, and, once finished, the system asks them whether they want to choose another skill, another test language, or exit.

As evident from the description above, DIALANG gives self-assessment a prominent position in the assessment procedure. This operationalises the belief that ability to self-assess is an important part of knowing and using a language. Every DIALANG user will be asked to self-assess their skills, and even if they choose to skip it, they will at least have been offered a chance to assess what they can do in the language that they are learning. If they do

complete the self-assessment section, the system compares their test performance and their self-assessment, and the match is reported as part of DIALANG feedback.

DIALANG test results and self-assessment are reported on the Council of Europe 6-level scale. The levels range from A1, Breakthrough, through A2, B1, B2, and C1, to Mastery at C2. The results include short verbal descriptions of what the levels mean in terms of ability to use the language. These descriptions also stem from the Council of Europe Common European Framework of Reference (Council of Europe 2001).

Initial development of the self-assessment section

The initial development of the self-assessment section was conducted by a self-assessment working group during DIALANG Phase 1 (for a report on the initial development of the assessment system, see Huhta *et al.*, forthcoming). The group members started the development from the scale descriptors available in Draft 2 of the Common European Framework, a working document brought out by the Council of Europe in 1996. An even larger set of descriptors was available in Brian North's PhD thesis (North 1995), and this was also used as a starting point for selecting potential statements for DIALANG self-assessment. The criteria for selecting the statements were that they should be concrete and practical enough for learners to understand; they should fit the general purpose orientation of the DIALANG test sections; and as a group, they should cover the whole range of ability from beginner to highly advanced. As a result, 36 potential self-assessment statements were identified for reading, 37 for writing, and 41 for listening. Three further statements were written for listening to complement skill levels that the working group felt were underrepresented.

The descriptors in the source documents had been developed for use by teachers and they were in a 'Can do' format. For self-assessment purposes, the members of the DIALANG self-assessment working group simplified the language in the statements a little, and changed the formulations from 'Can do' into 'I can'. For example, an original statement of 'Can follow speech which is very slow and carefully articulated, with long pauses for him/her to assimilate meaning' was changed into '*I can* follow speech which is very slow and carefully articulated, with long pauses *for me to get* the meaning'. Once the group had modified the statements, the DIALANG item-writing teams reviewed them and proposed some further modifications.

The revised set was vetted by Brian North for correspondence with the original statements, after which the statements were translated into the 13 other DIALANG languages in preparation for piloting. For each language, two translators translated the statements independently and, where there were discrepancies, they negotiated a joint solution. A multilingual working group then made cross-language comparisons to ensure the comparability of the

translations. After a final set of negotiations, the translated statements were implemented in the DIALANG Pilot Tool.

The initial development of the DIALANG self-assessment statements was driven by two conflicting content constraints. On the one hand, the statements should be worded so that they are easy to comprehend by learners, many of whom are not linguists. On the other hand, while the source statements had been written for teachers, they were also empirically linked to the Council of Europe scale, which had been adopted as the basis for the DIALANG system in 1997. Thus, while it might have been desirable to change the language of the source statements considerably and write some new statements with exact correspondences to the DIALANG task pool, it was deemed necessary to stay as close to the original statements as possible in order to avoid the extensive early validation work already conducted on the Council of Europe statements, for which the DIALANG project had neither the time nor the money. The result is a compromise which the project members hope is comprehensible to learners. Data from piloting and from early system use will be used to investigate whether this is the case.

Piloting – design and initial analyses

Two types of self-assessment were piloted in DIALANG: 'Overall' self-assessment, where learners read a six-level scale and select the level that best describes their own proficiency, and 'I can' statements, where learners read a set of detailed statements and answer Yes or No to each of them according to whether they can do what the statement says or not. The two formats for the listening skill are illustrated in Figures 1 and 2.

The two types of self-assessment were piloted alongside DIALANG test items with skill correspondence such that listening self-assessment statements were piloted with listening items, reading statements with reading items, and so on. Each participant was asked to give both an 'Overall' self-assessment and responses to 'I can' statements. Earlier, when DIALANG items in Finnish were piloted on paper, the participants were asked to answer all the 36–44 'I can' statements. In computer-based administration, in order to avoid fatigue, a matrix design was used so that each participant responded to 2/3 of the statements. This had an influence on the data available from the early part of DIALANG piloting, as will become evident in the report below.

In this article, I will use data and examples from the listening skill; the results for the other two skills were very similar. The data come from two test languages, Finnish and English, which are the two largest data sets available from the piloting for the time being. The sample size for the Finnish pilot was 451. Although the overall sample size for English was larger, 735 when this data set was extracted, the matrix design in data gathering meant that 223 respondents had answered listening self-assessment statements, and these

Figure 1 Listening 'Overall' self-assessment in the DIALANG Pilot Tool

You are now asked to assess your overall ability in one of the main language skills: reading, writing, speaking or listening. The computer has randomly assigned you the skill: **Listening**.

Please choose the statement below which most closely describes your ability in **listening** to the language which is being tested. If more than one of the statements is true, pick the 'best' one, i.e. the one nearest to the bottom of the screen. Press "Confirm" when you have given your response.

I can understand very simple phrases about myself, people I know and things around me when people speak slowly and clearly.	Yes
I can understand expressions and the most common words about things which are important to me, e.g. very basic personal and family information, shopping, my job. I can get the main point in short, clear, simple messages and announcements.	Yes
I can understand the main points of clear "standard" speech on familiar matters connected with work, school, leisure etc. In TV and radio current-affairs programmes or programmes of personal or professional interest, I can understand the main points provided the speech is relatively slow and clear.	Yes
I can understand longer stretches of speech and lectures and follow complex lines of argument provided the topic is reasonably familiar. I can understand most TV news and current affairs programmes.	Yes
I can understand spoken language even when it is not clearly structured and when ideas and thoughts are not expressed in an explicit way. I can understand television programmes and films without too much effort.	Yes
I understand any kind of spoken language, both when I hear it live and in the media. I also understand a native speaker who speaks fast if I have some time to get used to the accent.	Yes

constituted the data set reported here. Some of the groups compared in the report below are smaller depending on the criteria for grouping the data; the group sizes will be reported together with the results.

The background characteristics for the two participating samples, Finnish and English test takers, were similar in terms of their age range, gender distribution and educational background, but there was one potentially significant difference between them, namely self-assessed proficiency. The self-assessed ability levels of the participants in the Finnish pilots were fairly normally distributed across the ability range, while the ability distribution of the participants in the English pilots was slightly negatively skewed, i.e. the participants tended to assess themselves towards the higher end of the ability scale. This may have influenced the results reported below.

The self-assessment data can be analysed with respect to two questions that were interesting for the project in terms of test development, the relationship between the two types of self-assessment, i.e. 'Overall' scale and 'I can' statements, and the relationship between self-assessed proficiency and

Figure 2 Listening 'I Can' self-assessment in the DIALANG Pilot Tool

Below are a number of statements relating to your ability to **listen** to the tested language. Read each of the statements carefully and click

- "Yes" if you **CAN** do what the statement says and
- "No" if you **CANNOT** do what the statement says.

All questions must be answered. When you have finished, press "submit". Please make sure that you interpret each of the statements in relation to **listening** and not speaking, writing, reading, etc.

I can catch the main points in broadcasts on familiar topics of personal interest when the language is relatively slow and clear.	Yes / No
I can follow clear speech in everyday conversation, though in a real-life situation I will sometimes have to ask for repetition of particular words and phrases.	Yes / No
I can generally follow the main points of extended discussion around me, provided speech is clear and in standard language.	Yes / No
I can understand standard spoken language, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, academic or vocational life. Only extreme background noise, unclear structure and/or idiomatic usage causes some problems.	Yes / No
I can handle simple business in shops, post offices or banks.	Yes / No
I can follow specialised lecutres and presentations which use a high degree of colloquialism, regional usage or unfamiliar terminology.	Yes / No
I can understand questions and instructions and follow short, simple directions.	Yes / No

performance on DIALANG test items. The 'Overall' variable is categorical by nature, as it expresses which of the six ability categories each DIALANG user has chosen. In contrast, the summary variables for the 'I can' self-assessment and for performance on the DIALANG test items are continuous, in that they indicate how many 'Yes' responses to 'I can' statements and how many acceptable answers to test items each user has given. In the future, it will be possible to express these variables in terms of the Council of Europe scale. However, to make the translations in an accountable manner, there has to be a solid empirical basis for deciding which scores translate to which ability levels. Work is in progress in DIALANG to provide such standard setting rules (for a report on the early standard setting procedures, see Verhelst and Kaftandjieva 1999; for professional standards on standard setting, see AERA 1999, Chapter 4). For the time being, with the data sets as small as they are, the analyses reported here have been conducted on raw scores, i.e. numbers of 'Yes' responses to 'I can' statements and numbers of acceptable answers on the test items.

Figure 3 Relationship between Listening 'overall' and 'I can' self-assessment, Finnish

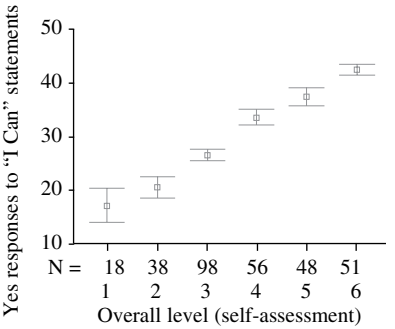
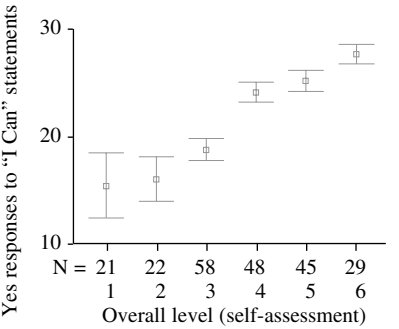


Figure 4 Relationship between Listening 'overall' and 'I can' self-assessment, English



Figures 3 and 4 give an overview of the relationship between the 'Overall' and 'I can' self-assessment in listening for Finnish and English. The values on the X-axis stand for the six 'Overall' categories from A1 to C2, while the Y-axis gives the number of Yes responses to the 'I can' statements. The centre of the box plot indicates the mean number of Yes responses within each category, and the whiskers show the 95% confidence interval for this value. The figures show a general rising tendency, which means that those who give themselves higher 'Overall' levels also tend to give more Yes-responses to the set of 'I can' statements presented to them.

The reader must be warned, however, that Figures 3 and 4 cannot be compared directly. The data-gathering design in the Finnish paper-based pilots was complete, i.e. the participants assessed themselves on the 'Overall' scale and answered all the 44 'I can' statements for listening. In the computer-based English pilots, the participants also assessed themselves on the 'Overall' scale, but they answered only 29 or 30 'I can' statements depending on the booklet that they got. Thus, while it may seem on a first glance that the 95% confidence interval for those who gave themselves the lowest level (the first box plot figure in each series) on the 'Overall' scale is wider for the participants in the English test, the absolute size of the interval is in fact very similar for the two groups. Given that the participants in the English test responded to fewer statements, the relative size of the interval is nevertheless slightly bigger.

The relationship between self-assessed overall ability level and performance on test items, depicted in Figures 5 and 6, also shows a rising tendency, which indicates a positive relationship between self-assessment and test performance. However, the figures also repeat the pattern from Figures 3 and 4 that the 95% confidence interval for the two lowest levels, A1 and A2, is wider than that for the rest of the levels. This may simply be the result of

Figure 5 Relationship between Listening 'overall' self-assessment and listening test performance, Finnish

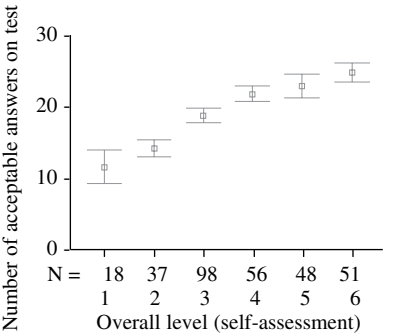
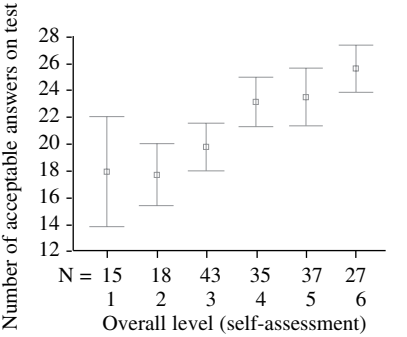


Figure 6 Relationship between Listening 'overall' self-assessment and listening test performance, English



small sample sizes, or it may reflect the fact that self-assessment at the beginning levels is not very accurate. However, it may also indicate that not all learners have the patience to read through the six levels of a self-assessment scale before deciding which level is appropriate for them, some may want to get on with the test and thus simply tick one of the first levels. Certainly, this explanation cannot be rejected out of hand on the basis of correlations between 'Overall' self-assessment, 'I can' statements, and test performance, as shown in Table 1.

Table 1a Correlations between self-assessment and test performance, Finnish

Skill: Listening	'Overall', scalar SA	'I can' statements	Raw score from test
'Overall', scalar SA	1.000 N=310	.826** N=309	.619** N=308
'I can' statements		1.000 N=315	.598** N=313
Raw score from test			1,000 N=451

**Correlation is significant at the 0.01 level (2-tailed).

Table 1b Correlations between self-assessment and test performance, English

Skill: Listening	'Overall', scalar SA	'I can' statements	Raw score from test
'Overall', scalar SA	1.000 N=223	.720** N=223	.416** N=175
'I can' statements	1.000	.489** N=223	 N=175
Raw score from test			1,000 N=175

**Correlation is significant at the 0.01 level (2-tailed).

All the correlations reported in Table 1 are statistically significant (p=.01). The values show that the relationship between the two types of self-assessment is closer than the relationship between either type of self-assessment and test performance. The differences between the correlations are also statistically significant (p=.000) with both Finnish and English. The correlations between either type of self-assessment and the test scores were of similar magnitude, i.e. the differences between the correlations were not statistically significant. Nevertheless, with the test of English, the absolute value of the correlation between the 'I can' statements and test performance was higher, and, while the difference was not statistically significant, this tendency was repeated in both languages with the other two skills. This issue thus needs to be investigated further. One possible way of gathering additional data might be to ask some pilot participants about their self-assessment strategies and their confidence in the assessments they gave either while they are taking a test or immediately afterwards.

The analyses that it has been possible to run so far are basic and the relatively small numbers mean that the results are only tentative. Comparisons are also made more difficult by the matrix design in data gathering and by the differences in data-gathering mode, one paper-based and the other computer-based. However, this is the type of data that early test development often has to rely on, and this is how we proceeded in DIALANG as well. Maximum sensible use was made of the available data when outlines for the upcoming system were developed, and possibilities were left open so that initial decisions could be modified in later versions of the system if deemed desirable.

Developments and revisions based on piloting

Comments from the participants in DIALANG pilots indicated that the number of 'I can' statements presented to the learners was too large, the statements began to look too alike and it was tiring to answer them all. From a practical acceptability point of view, it was therefore necessary to shorten this section. Moreover, based on early piloting data, it was not clear whether the 'Overall' self-assessment was effective and reliable, nor was it easy to develop a formula through which the two types of self-assessment could be combined for placement purposes or compared with test performance and reported back in a comprehensible way in the feedback section. It was therefore decided that the 'I can' section should be shortened, and that Version 1 of the DIALANG system should be published without 'Overall' self-assessment. However, while the shortened 'I can' section would be implemented in all further piloting, the 'Overall' self-assessment would remain a part of that design so that the connection between the two could be investigated further with a larger data set.

The criteria that the project set for the shortened 'I can' self-assessment section were that it should be informative at all the ability levels, as reliable as possible, and acceptable to DIALANG users. The means chosen to assist in the shortening decisions was a program called Optimal Test Design, developed by Angela Verschoor at CITOgroep. The program is based on a branch of mathematics called operations research, typical applications of which include train and aeroplane schedules. The program allows the user to set both statistical and content constraints before the optimisation algorithm is run. Possible content rules include 'at least N items from each of M categories', 'if A then not B', and 'at most 1 (or 2) of items X, Y, and Z'. This was important for DIALANG, because it gave us the opportunity to say that we wanted the shortened test to include at least two statements from each of the six Council of Europe skill levels, and it allowed us to identify groups of items that were worded in a very similar way and stipulate that only one or two of them should be included in the shortened set. The statistical constraints in Optimal Test Design are based on the test information function, and our criterion was that the shortened test should give an approximately equal amount of information over a broad range of ability levels, as broad as our original set of 'I can' statements allowed.

Based on comments from pilot participants, who had been given 'I can' sections of lengths ranging from 24 to 44 items, and on our judgements as test developers, together with trial analyses of test reliability for different lengths of test, we decided that the shortened test should have 18 statements for each skill. The reliabilities for the full 36–44 statement tests ranged from .93 to .95, while the reliabilities for the shortened tests were between .84 and .89. Though not ideal, we found this compromise between reliability and acceptability

satisfactory, as the placement decisions at the beginning of the test only needed to be approximate, and because high reliability could only be gained in exchange for a threat of user frustration or boredom, which would invalidate the results regardless of potential reliability.

The shortening analyses were run using all available data from DIALANG piloting until April 2001. After the analysis had been completed, the resulting tests were checked in content terms once more and one statement was exchanged for a near-equal but shorter and more broadly applicable one. The shortened Listening 'I can' test is reproduced in Appendix 1 as an example. The shortened versions are currently being piloted, and an added advantage of this is that the design is now complete, i.e. all test takers who get listening items in any DIALANG language will answer the same set of 18 'I can' statements. This will make comparisons across both test languages and administration languages more feasible.

Next steps in developing self-assessment in DIALANG

In addition to gathering more data on 'Overall' and 'I can' self-assessment, DIALANG is looking into expanding the self-assessment concept in computer-based assessment. One way that we see for this is learner self-rating. This direction was first explored in DIALANG during Phase I of the project, with experimental item types for writing and speaking. In the writing task, learners are asked to select a task and write a text based on a prompt on screen. Once complete, the task takes them to an evaluation screen, where they are guided to compare their own text against marked-up benchmarks. In the latter half of the year 2000, my colleague Mirja Tarnanen and I ran an experiment and conducted a usability study of this application with learners of Finnish as a foreign language, on which we reported at two conferences in 2001 (Luoma and Tarnanen 2001a, 2001b). The results indicated that the learners found the task educative and interesting, and they were able to use the marked-up benchmarks to assess their own proficiency. This encouraging feedback led to plans to extend the experiment into English, French, and German, a study of which is currently under way.

Conclusion

The development of the self-assessment strand in DIALANG continues, and with larger data sets it will be possible to investigate differences between languages and possibly regions of origin, i.e. whether learners from some regions of the world tend to over- or underestimate their skills, as has been suggested (e.g. Oscarson 1997). It is important to remember, however, that even if data are available, the meaning is not necessarily apparent from a mere inspection of numbers. Additional insights need to be gained from learners

using the system. Scale descriptors, for instance, are meaningful and important for teachers and applied linguists, but learner interpretations of the descriptors and differences between different descriptors are equally important if we are to achieve a truly learner-orientated assessment system. Decisions in DIALANG about whether or not to include 'Overall' self-assessment in the system must likewise be informed not only by statistical and programming considerations, important as these are, but also by learner feedback on how they interpret the scale and whether they find it useful. Moreover, if statistically significant differences are found between self-assessments of learners of Spanish and learners of Dutch, for example, we must carefully search for content interpretations for the differences. This is why our project members are conducting research into how learners interact with the DIALANG system (Figueras and Huhta, in progress).

The prominence given to self-assessment in DIALANG is based on the belief that it is beneficial for learners and that it promotes learner independence. Since the whole DIALANG system is learner-orientated, it also fits the system ideologically. However, we do not expect great changes in learner orientation 'as a result of being exposed to DIALANG', nor are we planning any experimental pre-DIALANG, post-DIALANG designs to detect this. Rather, we expect the effects to be subtle. Being asked to assess one's language ability raises awareness in DIALANG users that such evaluations can be made, and through this, DIALANG brings its contribution to the array of concepts that language learners have for evaluating their own proficiency. The basic DIALANG system complements the self-assessment by comparing it with test results and providing the users with information about why the two might differ. Preliminary feedback from our experiment with self-rating indicates that while learners feel able to assess their own proficiency, they also need external assessment to form a picture of their skills that they can rely on. Self-assessment is therefore not the be-all and end-all of all assessment, but we would like to believe that it is a useful addition to the assessment arsenal that the modern world offers language learners.

References

AERA 1999. [American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999.] *Standards for Educational and Psychological Testing*. Washington, D.C.: AERA.

Council of Europe 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: CUP.

Figueras, N. and A. Huhta (in progress). Investigation into learner reception of DIALANG.

Huhta, A., S. Luoma, M. Oscarson, K. Sajavaara, S. Takala, and A. Teasdale (forthcoming). DIALANG – a diagnostic language assessment system for learners. In J. C. Alderson (ed.) Case studies of the use of the Common European Framework. Council of Europe.

Luoma, S. and M. Tarnanen. 2001a. Experimenting with self-directed assessment of writing on computer – Part I: self-assessment versus external assessment. Paper given in a symposium entitled 'Learning-centred assessment using information technology' at the annual Language Testing Research Colloquium in St Louis, Missouri February 20–24, 2001.

Luoma, S. and M. Tarnanen. 2001b. Experimenting with self-directed assessment of writing on computer – Part II: learner reactions and learner interpretations. Paper given at the annual conference of the American Association of Applied Linguistics in St Louis, Missouri February 24–27, 2001.

North, B. 1995. *The development of a common framework scale of language proficiency based on a theory of measurement*. PhD thesis, Thames Valley University, November 1995.

Oscarson, M. 1997. Self-assessment of foreign and second language proficiency. In Clapham, C. and D. Corson (eds.), *Encyclopedia of Language and Education, Volume 7: Language testing and assessment*. 175–187. Dordrecht: Kluwer.

Verhelst, N. D. and F. Kaftandjieva. 1999. A rational method to determine cutoff scores. Research Report 99-07, Faculty of Educational Science and Technology, Department of Educational Measurement and Data Analysis, University of Twente, The Netherlands.

Appendix 1

Shortened DIALANG 'I can' self-assessment section for listening

CEF level	DIALANG Listening 'I can' statement
A1	I can follow speech which is very slow and carefully articulated, with long pauses for me to get the meaning.
A1	I can understand questions and instructions and follow short, simple directions.
A2	I can understand enough to manage simple, routine exchanges without too much effort.
A2	I can generally identify the topic of discussion around me which is conducted slowly and clearly.
A2	I can understand enough to be able to meet concrete needs in everyday life provided speech is clear and slow.
A2	I can handle simple business in shops, post offices or banks.
B1	I can generally follow the main points of extended discussion around me, provided speech is clear and in standard language.
B1	I can follow clear speech in everyday conversation, though in a real-life situation I will sometimes have to ask for repetition of particular words and phrases.
B1	I can understand straightforward factual information about common everyday or job-related topics, identifying both general messages and specific details, provided speech is clear and a generally familiar accent is used.
B1	I can catch the main points in broadcasts on familiar topics and topics of personal interest when the language is relatively slow and clear.
B2	I can understand standard spoken language, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, academic or vocational life. Only extreme background noise, unclear structure and/or idiomatic usage causes some problems.
B2	I can follow the essentials of lectures, talks and reports and other forms of presentation which use complex ideas and language.
B2	I can understand announcements and messages on concrete and abstract topics spoken in standard language at normal speed.
B2	I can understand most radio documentaries and most other recorded or broadcast audio material delivered in standard language and can identify the speaker's mood, tone, etc.
C1	I can keep up with an animated conversation between native speakers.
C1	I can follow most lectures, discussions and debates with relative ease.
C2	I have no difficulty in understanding any kind of spoken language, whether live or broadcast, delivered at fast native speed.
C2	I can follow specialised lectures and presentations which use a high degree of colloquialism, regional usage or unfamiliar terminology.

Section 3

A European View

9 Council of Europe language policy and the promotion of plurilingualism

Joseph Sheils
Modern Languages Division/Division des langues vivantes,
DGIV Council of Europe/Conseil de l'Europe, Strasbourg

Introduction

I am grateful to ALTE for taking the initiative in organising this significant event in the European Year of Languages calendar and for the opportunity to present some aspects of the Council of Europe's work in promoting plurilingualism in the context of the Year. The aims of the Year, which is jointly organised by the Council of Europe and the European Union, can be summarised as: to encourage more people to learn (more) languages and to raise awareness of the importance of protecting and promoting the rich linguistic diversity of Europe. The main messages of the Year are captured in two slogans: 'Languages open doors' – to opportunities, social inclusion, tolerance of differences; and 'Europe, a wealth of languages' – the more than 200 indigenous languages in the 43 member states of the Council of Europe and the languages of migrant communities are equally valid as modes of expression for those who use them.

In a Europe which is and will remain multilingual, policy responses to this reality lie between two ends of a continuum. There is on the one hand policy for the reduction of diversity, and on the other for the promotion and maintenance of diversity; both can be pursued in the name of improved potential for international mobility, improved communication and economic development. The Council of Europe has always pursued the important aim of maintaining the European cultural heritage, of which linguistic diversity is a significant constituent, and for which linguistic diversity provides the vital conditions. It does so through legal instruments such as the European Charter for Regional or Minority Languages, which continues to be ratified by an increasing number of states, and through its programmes of intergovernmental co-operation involving the 48 states party to the European Cultural

Convention. Its technical instruments play a major role in promoting linguistic diversity, in particular the 'Threshold Level' and related level descriptions developed for over 30 national and regional languages, the Common European Framework of Reference for Languages, and more recently the European Language Portfolio.

Linguistic diversity and plurilingualism

'Linguistic diversity' has to be considered under two concepts: 'multilingualism' and 'plurilingualism':

We are currently developing a draft guide to language education policy development, in which we define these as follows (1):

> 'multilingualism' refers to the presence in a geographical area, large or small, of more than one 'variety of language,' i.e. the mode of speaking of a social group whether it is formally recognised as a language or not; in such an area individuals may be monolingual speaking only their own variety.
>
> 'plurilingualism' refers to the repertoire of varieties of language that many individuals use, and is therefore the opposite of monolingualism; it includes the language variety referred to as 'mother tongue' or 'first language' and any number of other languages or varieties.

Thus in some multilingual areas some individuals are monolingual and some are plurilingual.

The Council of Europe's aims are to promote the development of language education policies that enable all Europeans to become plurilingual in ways that are appropriate to the area where they live. Language learning in school is the beginning of this process, but adult education has a key role to play because the development of plurilingualism is essentially a lifelong activity.

We are guided by a number of basic principles, including the following as set out in the draft guide on language education policy:

> *individual plurilingualism is a significant influence on the evolution of a European identity*: as Europe is a multilingual area in its entirety and, in any given part, the sense of belonging to Europe and the acceptance of a European identity is dependent on the ability to interact and communicate with other Europeans using the full range of one's linguistic repertoire
>
> *plurilingualism is plural*: because of the variation in multilingualism in different parts of Europe, the plurilingualism of individuals has to be appropriate to the area where they live; there is no preferred or recommended model of plurilingualism and the plurilingualism of the individual may change with mobility and throughout lifelong learning
>
> *the development of plurilingualism is possible*: it is a question of access for all who so wish and it is also a question of appropriate objectives and methods; the Council of Europe's *Common European Framework of*

> *Reference for Languages* and other technical instruments such as the European Language Portfolio already provide a valuable basis; the opportunity to develop one's plurilingualism should and can be made available to all, both in education systems and as part of lifelong learning.

The implications of plurilingualism for language teaching

In recent years, the concept of plurilingualism has grown in importance in the Council of Europe's approach to language learning. It is central to the *Common European Framework of Reference* (2):

> 'The plurilingual approach emphasises the fact that as an individual person's experience of language in its cultural contexts expands, from the language of the home to that of society at large and then to the languages of other peoples (whether learnt at school or college, or by direct experience), he or she does not keep these languages and cultures in strictly separated mental compartments. Rather, the learner builds up a communicative competence to which *all* knowledge and experience of language contributes and in which languages interrelate and interact. In different situations, a person can call flexibly upon different parts of this competence to achieve effective communication with a particular interlocutor.'

Accordingly, the aim is 'no longer seen as simply to achieve 'mastery' of one or two, or even three, languages, each taken in isolation, with the 'ideal native speaker' as the ultimate model. Instead, the aim is to develop a linguistic repertory, in which all linguistic abilities have a place.'

As language learning is a lifelong task, the development of a young person's motivation, skill and confidence in facing new language experience out of school comes to be of central importance. Schools must prepare young people for the lifelong development of plurilingualism and adult education providers have the challenge of ensuring coherence in the continuation of this process and in further extending it. Accordingly (to paraphrase the Framework and John Trim), the responsibilities of individual institutions and their teachers cannot simply be confined to the attainment of a given level of proficiency in a particular language at a particular moment in time, important though that undoubtedly is.

The recent developments in the Council of Europe's language programme have been designed to produce tools for the language profession and for learners to facilitate the coherent promotion of plurilingualism. In particular, The European Language Portfolio provides a personal document for individuals of all ages in which they can show their competences in different languages in an internationally transparent manner and show their significant contacts with other cultures. Learning experiences of all kinds and at all levels can be given recognition. In addition, the *Common European Framework of*

Reference for Languages provides language professionals with a means, inter alia, of specifying objectives and assessing achievements.

The development and widespread dissemination of these instruments has been recommended by the Committee of Ministers and the Standing Conference of European Ministers of Education (Cracow October 2000) and this is a priority for the Council of Europe in the European Year of Languages 2001 (3).

The Common European Framework

As the contents of the Framework have been explained in some detail elsewhere (Trim 2002), I shall simply summarise certain elements by way of an introduction to the European Language Portfolio scheme.

The Framework can facilitate the mutual recognition of language qualifications and also the promotion of diversification, learner independence and plurilingualism. It offers a descriptive scheme that is very valuable for planning language teaching, particularly for setting objectives. It is equally valuable for assessment, testing and qualification. Clearly, for reasons of comparability and mobility, different systems of certification need to be related on a common framework that is widely accepted in Europe – and even beyond. The Council of Europe Framework was designed with this in mind.

As illustrated below (Figure 1), the scheme has three broad bands defined with six levels which can be subdivided into finer levels as required. A distinction is made between interaction and production in speaking and writing because these involve different skills. The descriptors are framed in positive terms – stating what the learners can do – rather than what they cannot do, and how well rather than how badly they can perform.

Figure 1 Common European framework of reference for languages
Cadre Europeen Commun de reference pour les langues

- Proficiency levels/*Niveau de référence*:
  - basic user/*utilisateur élémentaire* A1A2
  - independent user/*utilisateur indépendant* B1B2
  - proficient user/*utilisateur experimenté* C1C2
- Proficiency dimensions/*Compétences*:
  - understanding: listening, reading/*comprendre: écouter, lire*
  - speaking: spoken interaction, spoken production/*parler: prendre part à une conversation, s'exprimer oralement en continu*
  - writing: written interaction, written production/*écrire: (interaction, production)*
- Proficiency descriptors/*Descripteurs*:
  - positive 'can do' statements/*je peux…' (descripteurs formulés en termes positifs)*

The self-assessment grid from the Framework is a core element in all accredited versions of a European Language Portfolio (Appendix 1). It

distinguishes the traditional four skills but separates monologue (spoken production) from conversation (spoken interaction).

It should be noted that the ALTE can-do statements and the specifications and scales in the DIALANG scheme are contained in the appendices of the Common European Framework book.

The Framework is now published by CUP and Editions Didier. It is also available in downloadable form on the Council of Europe portfolio website and we are grateful to the publishers for their generous co-operation in this regard: http://culture.coe.int/portfolio.[1]

I wish to acknowledge the contribution of the authors: J. Trim, B. North and D. Coste, and also of the Swiss National Research Council Project, which provided the basis for the scales of descriptors in the Framework.

The European Language Portfolio

The Council of Europe has developed a related instrument – the European Language Portfolio (ELP), which can facilitate the implementation in practice of the goals, standards and methodological options described in the Framework.

The Committee of Ministers recommended to the governments of member states in 1998:

> RECOMMENDATION No. R (98) 6 OF THE COMMITTEE OF MINISTERS TO MEMBER STATES CONCERNING MODERN LANGUAGES
>
> 27. Encourage the development and use by learners in all educational sectors of a personal document (European Language Portfolio) in which they can record their qualifications and other significant linguistic and cultural experiences in an internationally transparent manner, thus motivating learners and acknowledging their efforts to extend and diversify their language learning at all levels in a lifelong perspective.

(The full text is available at http://cm.coe.int/site2/ref/dynamic/-recommendations.asp)

The Modern Languages Division in Strasbourg initiated a pilot scheme to design and pilot ELP models and to explore how they have worked in the daily life of education institutions over the last three years. Some 30,000 learners were involved from primary, secondary, vocational, higher and adult education in 15 countries. Some 700 were learners in adult education taking part mainly through the EAQUALS network, which has since co-operated

1 Key documents and information are available on the Council of Europe Portfolio website: http://culture.coe.int/portfolio. The text of the Common European Framework of Reference is also available on this site. Further information can be obtained from: Modern Languages Division, DG IV, Council of Europe, 67075 Strasbourg, France.

with ALTE to produce a joint European Language Portfolio. A report on the pilot scheme and other key documentation is available on the portfolio website indicated above.

The Portfolio scheme is now being widely introduced in the Council of Europe's member states and introductory seminars are being organised for nominated representatives of all countries.

What is a European Language Portfolio?

It is a personal document that belongs to the learner. It can take different forms depending on the age of learners. Three basic formats have been developed to suit the needs of young learners, secondary school/adolescent learners and adults. While the ELP respects the diversity of learning contexts in our member states, all Portfolios are being developed according to certain agreed Principles and Guidelines. A certain unity in diversity and quality assurance is ensured by the fact that each ELP has to be approved by a Council of Europe Validation Committee, which grants the right to use the special Council of Europe logo.

All ELPs have three components:

- Language Passport
- Language Biography
- Dossier

The *Principles* and *Guidelines* approved by the Council of Europe define the three components of the ELP as follows (4):

- The Language Passport section provides an overview of the individual's proficiency in different languages at a given point in time; the overview is defined in terms of skills and the common reference levels in the Common European Framework; it records formal qualifications and describes language competencies and significant language and intercultural learning experiences; it includes information on partial and specific competence; it allows for self-assessment, teacher assessment and assessment by educational institutions and examinations boards.
- To facilitate pan-European recognition and mobility, a standard presentation of a Language Passport is promoted by the Council of Europe for ELPs for adults.
- The Language Biography facilitates the learner's involvement in planning, reflecting upon and assessing his or her learning process and progress; it encourages the learner to state what he/she can do in each language and to include information on linguistic and cultural experiences gained in and outside formal educational contexts; the biography varies according to the learning contexts but is organised to promote plurilingualism, i.e. the development of competencies in a number of languages.

- The Dossier offers the learner the opportunity to select relevant and up-to-date materials to document and illustrate achievements or experiences recorded in the Language Biography or Language Passport.

What is the ELP's function?

The ELP has two functions: reporting and pedagogic. These are defined and illustrated in the *Guide for Developers of a European Language Portfolio* (Günther Schneider and Peter Lenz) and the *Guide for Teachers and Teacher Trainers* (David Little and Radka Perclová). Much of what follows is directly taken from or based on these two publications (5).

> *Reporting*. The European Language Portfolio aims to document its holder's plurilingual language proficiency and experiences in other languages in a comprehensive, informative and transparent way. The instruments contained in the ELP help learners to take stock of the levels of competence they have reached in their learning of one or several languages in order to enable them to inform others in a detailed and internationally comparable manner.

There are many occasions when it can be useful to present a Language Portfolio that is up to date, e.g. a transfer to another school, change to a higher educational sector, the beginning of a language course, a meeting with a career advisor, or an application for a new post. In these cases the ELP is addressed to persons who have a role in decisions that are important for the owner of the Language Portfolio. Those who receive an ELP from a learner for qualification and similar purposes may be particularly interested in:

- the results of relevant summative evaluation
- diplomas and certificates
- other 'evidence', which may not normally be included in certification, e.g.
  - up-to-date information based on self-assessment
  - participation in exchange programmes, practical training in other language regions
  - attestations/descriptions of regular private contacts with speakers of other languages, professional correspondence in other languages
  - information on objectives, curricula, examination criteria, etc.
  - selected written products, audio and video recordings of oral productions, etc.

'Evidence' of this kind is important because successful language learning cannot always be documented by means of exams or diplomas. For example:

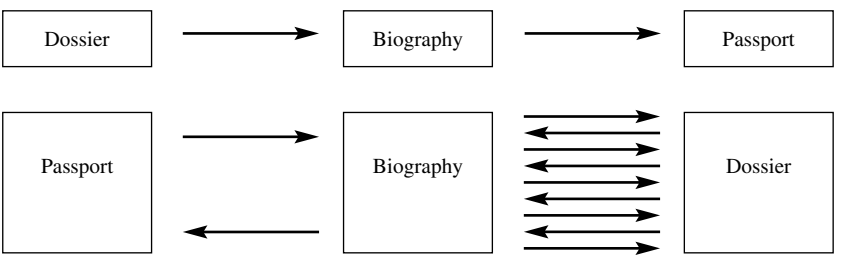
- a learner may improve his/her language skills considerably through stays abroad *after* receiving a diploma;
- some aspects such as intercultural competence are rarely tested through exams. For some languages no standardised exams or certificates exist (in many cases languages of migrants and their children).

The purpose of the ELP is not to replace the certificates and diplomas that are awarded on the basis of formal examinations, but to supplement them by presenting additional information about the owner's experience and concrete evidence of his or her foreign language achievements. Clearly, the importance of the ELP's reporting function will vary according to the age of the owner. It will usually be much less important for learners in the earlier stages of schooling than for those approaching the end of formal education or already in employment. For this reason the Council of Europe has introduced a standard Language Passport for adults only. It is particularly important to adult learners that the ELP should be accepted internationally, and this is more likely to happen if the first of its components is the same everywhere.

- *Pedagogical*. The ELP is also intended to be used as a means of making the language-learning process more transparent to learners, helping them to develop their capacity for reflection and self-assessment, and thus enabling them gradually to assume increasing responsibility for their own learning. This function coincides with the Council of Europe's interest in fostering the development of learner autonomy and promoting lifelong learning.

These two functions are closely linked in practice as illustrated in the *ELP Guide for Teachers and Teacher Trainers*.

Figure 2 European language portfolio



European Language Portfolio
Guide for Teachers and Teacher Trainers
(D. Little and R. Perclová)

For example, with younger learners the teacher might start with the Dossier, in which learners are encouraged to keep examples of their best work. At a somewhat later stage the Biography is introduced and learners are helped to set their own learning targets and to review their learning progress. At a still later stage the Passport is introduced so that learners can take stock of their developing linguistic identity using the self-assessment grid from the Common European Framework.

The process can be reversed and this may suit older learners. The Language Passport might be introduced at the beginning as a way of challenging learners to reflect on their linguistic identity and the degree of proficiency they have already achieved in their target language or languages. They then proceed to the Biography and set individual learning targets. Learning outcomes are collected in the Dossier and evaluated in the Biography, and this provides the basis for setting new goals. The process is repeated until the end of the course, when the learners return to the Passport and update their self-assessment.

Clearly the emphasis placed on the different functions of the Portfolio may vary depending on the nature and length of courses, but both functions are important and all three parts must be included in any European Language Portfolio if it is to be accredited by the Council of Europe. Draft models can be submitted to the European Validation Committee for accreditation. Details can be obtained from the Modern Languages Division in Strasbourg and all relevant documents, including *Rules for Accreditation*, are available on the portfolio website (see footnote on page 163).

The Portfolio and assessment

The successful integration of the ELP with programmes and classroom practice depends partly on establishing the right kind of relation between self-assessment and the various forms of assessment by others – test and examinations – to which learners are subject.

The *Guide for Teachers and Teacher Trainers* (Little and Perclová) links summative and formative self-assessment:

> 'When learners assess themselves in the passport component, they are engaging in a form of summative assessment: a statement of their proficiency at a particular point in their lives. On the other hand, the self-assessment that runs through the biography component and keeps the contents of the dossier under critical review has a formative function; and it is so thoroughly integrated with the pedagogical purpose of the ELP that it is as much a habit of mind as an activity. However, these two kinds of self-assessment depend on the same complex of knowledge, self-knowledge and skills, which means that learners are likely to be more proficient in performing summative self-assessment if formative self-assessment – what one might call reflective self-evaluation – has been an integral part of their learning experience' (p. 55).

The *Guide* points out that self-assessment is concerned with three basic elements of the learning process itself: learners need to be able to assess how well they are progressing overall, how well they are learning at a particular stage, and how successful they are in performing individual learning tasks and meeting specific learning goals. Self-assessment with this focus is an integral part of the reflective approach to learning. It is equally concerned with the learner's proficiency in terms of the Council of Europe's scales and

descriptors and with the learner's linguistic competence.

It should be stressed again that the self-assessment required by the ELP is not intended to replace assessment of learners by teachers, schools, universities or public examination boards: the language passport contains the owner's assessment of his or her foreign language proficiency, but it also provides space in which to record examinations passed and certificates awarded. Ideally, of course, self-assessment and assessment by others should complement each other. Self-assessment is based on the learner's developed capacity to reflect on his or her own knowledge, skills and achievement, while assessment by others provides an external, objective measure of the same knowledge, skills and achievements.

Conclusion: the contribution of the ELP to Council of Europe goals

The ELP should be seen in the context of the Council of Europe's programmes intended to value and promote linguistic diversity and to support individuals in diversifying their language learning. It complements other initiatives to deepen mutual understanding and tolerance in our multilingual and multicultural societies. Moreover, by developing responsibility and self-reliance in learners with regard to decisions about learning, it can support young people in developing the more general competences necessary for socially responsible participation in processes relating to active democratic citizenship. The Portfolio has the potential to ensure coherence in the transition between education sectors and in lifelong language learning. It can facilitate mobility as it provides an internationally transparent record of competences. The Standing Conference of European Ministers of Education has recommended that governments of member states '*take steps to ensure that the ELP is acknowledged as a valid record of competences regardless of its country, region, sector or institution of origin*'.

The ELP is an instrument with which to motivate learners and to support them in developing plurilingualism throughout life. It will enable all the partners at successive stages and in different contexts of language learning to contribute in a coherent and transparent manner to the development of an individual's plurilingualism.

Wherever appropriate, Council of Europe member states are currently setting up a national instance to co-ordinate the ELP in compulsory education at a national level. This body should help to develop a co-ordinated policy with all the partners. It will also act as a pre-validation agency, forwarding draft ELPs for the school sector with a Recommendation to the European Validation Committee in Strasbourg. It is important to examine the potential role of international NGOs and associations in this development and

dissemination process in the field of adult education and training.

In order to promote plurilingualism, education systems formal and informal, obligatory and post-obligatory, need to provide forms and organisations of language learning that promote an integrated competence and a consciousness of learners' existing repertoires and of their potential to develop and adapt those repertoires to changing circumstances.

The draft guide for language education policy describes this plurilingualism in its various guises as characterised by:

- a repertoire of languages and language varieties
- competences of different kinds and levels within the repertoire
- an awareness of and an ability to use transferable skills in language learning
- a respect for the plurilingualism of others and for the significance of languages and varieties irrespective of their social status
- a respect for the cultures embodied in languages and for the cultural identities of others
- an awareness of the potential for participation in democratic and other social processes in multilingual societies afforded by their plurilingualism.

The Council of Europe is grateful for the contribution of ALTE to various aspects of its activities in promoting plurilingualism, including earlier work on *Vantage Level* and more recently *Breakthrough*, and its involvement in the European Language Portfolio scheme with EAQUALS. We warmly welcome their joint ELP, which is one of the first to be validated by the European Validation Committee.

The adult education sector, in all its diversity, is a key partner in the lifelong process of helping and supporting individuals to develop plurilingualism. Providers must build on existing competences and previous learning experiences in responding to the specific needs of adults. The challenge, in an essentially market-driven sector, is to show the added value of the European Language Portfolio in ensuring coherence and transparency in adult education provision. The Council of Europe looks forward to the wider involvement of adult education bodies and institutions throughout Europe in responding to this challenge.

Appendix 1

		A1	A2	B1	B2	C1	C2
Understanding	Listening	I can understand familiar words and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly.	I can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevance (e.g. very basic personal and family information, shopping, local area, employment). I can catch the main point in short, clear, simple messages and announcements.	I can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, etc. I can understand the main point of many radio or TV programmes on current affairs or topics of personal or professional interest when the delivery is relatively slow and clear.	I can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. I can understand most TV news and current affairs programmes. I can understand the majority of films in standard dialect.	I can understand extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly. I can understand television programmes and films without too much effort.	I have no difficulty in understanding any kind of spoken language, whether live or broadcast, even when delivered at fast native speed, provided I have some time to get familiar with the accent.
	Reading	I can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues.	I can read very short, simple texts. I can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus and timetables and I can understand short simple personal letters.	I can understand texts that consist mainly of high frequency everyday or job-related language. I can understand the description of events, feelings and wishes in personal letters.	I can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. I can understand contemporary literary prose.	I can understand long and complex factual and literary texts, appreciating distinctions of style. I can understand specialised articles and longer technical instructions, even when they do not relate to my field.	I can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.
Speaking	Spoken Interaction	I can interact in a simple way provided the other person is prepared to repeat or rephrase things at a slower rate of speech and help me formulate what I'm trying to say. I can ask and answer simple questions in areas of immediate need or on very familiar topics.	I can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar topics and activities. I can handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself.	I can deal with most situations likely to arise whilst travelling in an area where the language is spoken. I can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events).	I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible. I can take an active part in discussion in familiar contexts, accounting for and sustaining my views.	I can express myself fluently and spontaneously without much obvious searching for expressions. I can use language flexibly and effectively for social and professional purposes. I can formulate ideas and opinions with precision and relate my contribution skilfully to those of other speakers.	I can take part effortlessly in any conversation or discussion and have a good familiarity with idiomatic expressions and colloquialisms. I can express myself fluently and convey finer shades of meaning precisely. If I do have a problem I can backtrack and restructure around the difficulty so smoothly that other people are hardly aware of it.
	Spoken Production	I can use simple phrases and sentences to describe where I live and people I know.	I can use a series of phrases and sentences to describe in simple terms my family and other people, living conditions, my educational background and my present or most recent job.	I can connect phrases in a simple way in order to describe experiences and events, my dreams, hopes and ambitions. I can briefly give reasons and explanations for opinions and plans. I can narrate a story or relate the plot of a book or film and describe my reactions.	I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue, giving the advantages and disadvantages of various options.	I can present clear, detailed descriptions of complex subjects integrating sub-themes, developing particular points and rounding off with an appropriate conclusion.	I can present a clear, smoothly flowing description or argument in a style appropriate to the context and with an effective logical structure that helps the recipient to notice and remember significant points.
Writing	Writing	I can write a short, simple postcard, for example sending holiday greetings. I can fill in forms with personal details, for example entering my name, nationality and address on a hotel registration form.	I can write short, simple notes and messages. I can write a very simple personal letter, for example thanking someone for something.	I can write simple connected text on topics which are familiar or of personal interest. I can write personal letters describing experiences and impressions.	I can write clear, detailed text on a wide range of subjects related to my interests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. I can write letters highlighting the personal significance of events and experiences.	I can express myself in clear, well-structured text, expressing points of view at some length. I can write about complex subjects in a letter, an essay or a report, underlining what I consider to be the salient issues. I can select a style appropriate to the reader in mind.	I can write clear, smoothly flowing text in an appropriate style. I can write complex letters, reports or articles that present a case with an effective logical structure which helps the recipient to notice and remember significant points. I can write summaries and reviews of professional or literary works.

References

Beacco, J-C. and M. Byram (unpublished draft). *Guide for the Development of Language Education Policies in Europe* (Provisional title. Publication of draft 1 in spring 2002.)

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press, pp. 4–5 (French edition: Editions Didier 2001).

Council of Europe. 2000. *Education Policies for Democratic Citizenship and Social Cohesion: Challenges and Strategies for Europe. Adopted texts*. Cracow (Poland) 15–17. October 2000, 21–24.

Council of Europe – Education Committee. 2000. European Language Portfolio: Principles and Guidelines. Document DGIV/EDU/LANG (2000) 33 (see also Portfolio website).

Little, D. and R. Perclová. 2001. *European Language Portfolio. Guide for Teachers and Teacher Trainers*. Strasbourg: Council of Europe.

Schneider, G. and P. Lenz. 2001. *Guide for Developers of a European Language Portfolio*. Strasbourg: Council of Europe.

Trim, J. 2002. Understanding the Common European Framework of Reference for Languages: Learning, Teaching, Assessment Conference (CEF). Paper given at 25th TESOL – Spain Convention, Madrid.

10 Higher education and language policy in the European Union

Wolfgang Mackiewicz
Conseil Européen pour les Langues/
European Language Council

Introduction

In this paper, I shall be dealing with EU language and language education policy and about policy development undertaken by the Conseil Européen pour les Langues/European Language Council and the Thematic Network in the Area of Languages in response to linguistic challenges posed by European integration. I shall conclude by reflecting on measures that universities need to take if they are to live up to their new role as institutions belonging to a European area of higher education. In doing so, I shall have occasion to refer to the draft Berlin Declaration issued by the members of the Scientific Committee of the Berlin European Year of Languages 2001 Conference held at the Freie Universität Berlin on 28–30 June 2001. In addition, I shall repeatedly refer to the contributions made by the Council of Europe to the promotion of multilingualism in Europe.

In discussing EU language policy, I shall be considering language policy at Community level – not language policy at member state or regional level. However, during the course of the paper, I shall refer to what I regard as the responsibility and duties of the member states and the regions in regard to language education policy.

EU language policy

As regards EU language policy, two things need to be noted above all else. (i) In education, the powers of the EU are limited. This is evidenced by the fact that Article 149 of the Treaty Establishing the European Community describes Community action in the fields of education, vocational training and youth in terms of contributing to, supporting and supplementing action taken by the member states. The Council of the European Union can (jointly with the

European Parliament) adopt incentive measures, for example action programmes such as Socrates, and it can, on a proposal from the Commission, adopt recommendations, such as the Council Resolution of 16 December 1997 on the early teaching of European Union languages. Council decisions in the field of education take account of the respective powers at each tier of authority; they may have the status of secondary legislation. The principle of subsidiarity applies. (ii) There is not one single Community document that sums up what might be called EU language policy. There are a substantial number of documents that are either devoted to a specific aspect of language policy or language education policy or that mention language policy issues in a wider context: the Treaty, Council decisions and resolutions, documents published by the European Commission (White Papers, Green Papers, Communications, etc.), resolutions of the European Parliament, and opinions of the Social and Economic Committee and the Committee of the Regions, etc. These documents bear witness to the evolvement of EU language policy over time; they also reveal subtle differences of opinion between the different institutions. It is an indication of this state of affairs that the Council asked the Commission in 1995 'to make an inventory of the Community's policies and activities which take linguistic diversity and multilingualism into account, and to assess those policies and activities'.

If one takes all the relevant documents together, one can identify a number of principles which constitute the core of EU language policy. The following documents are particularly relevant in this respect: the Treaty Establishing the European Community, i.e. the Treaty of Amsterdam, the Council conclusions of 12 June 1995 on linguistic and cultural diversity and multilingualism in the European Union, the Council Resolution of 31 March 1995 on improving and diversifying language learning and teaching within the education systems of the European Union, the Commission's White Paper *Teaching and Learning: Towards the Learning Society* of 1995 and the decisions establishing the Community's action programmes in the fields of education and training. Among the documents that are particularly relevant to my theme is the Commission's *Memorandum on Higher Education in the European Community* of 1991, which under the general heading of 'The European Community Dimension' has a section entitled 'The Central Importance of Language'.

Linguistic diversity

The fundamental principle underlying all of the EU's statements on matters regarding language policy is that of linguistic and cultural diversity. Unlike the United States of America, the European Union is being constructed as a multilingual society. The eleven national languages of the member states are regarded as being equal in value. (In addition, Irish enjoys a special status;

Letzeburgesch does so to a lesser degree. Both, of course, have the official status of national language in their respective countries.) The promotion of linguistic diversity is at the very heart of the unique experience of European integration. The idea of a common language – a Union *lingua franca* – is rejected.

Linguistic diversity is regarded as important for political as well as for cultural reasons. For political reasons, because the principle whereby Union citizens enjoy equal status and equal rights includes the right to one's own language. For cultural reasons, because the different languages are seen as vehicles of the wealth of European culture, which is one of the characteristic features of European society and must remain so.

The most visible sign of the political dimension of linguistic diversity is the fact that all Community legislation – the so-called *acquis communautaire* – is translated into all the official Community languages. Translating the *acquis* is a major task. For example, when Finland joined the Union in 1995, the translation of the *acquis* had not been completed. In order to avoid a repetition of this, the *acquis* is currently being translated into the languages of the candidate countries that are expected to join the Union in the foreseeable future. Translation is complemented by interpreting. In the European Parliament there is interpreting from every Community language into the other ten Community languages. Interpreting also plays a major role at Council meetings and at meetings arranged by the European Commission. However, there are limits to this. For example, at the informal meetings of the Councils of Ministers, interpreting is only provided for English, French, and the language of the member state that holds the presidency – a procedure that was last challenged by the German government at the beginning of the Finnish presidency in July 1999.

The principle of linguistic diversity also applies to contacts between the citizen and the European institutions and bodies carrying out the tasks entrusted to the Community, such as the European Parliament, the European Commission, and the Ombudsman. Article 21 of the Treaty stipulates that 'every citizen may write to any of the (European) institutions or bodies … in one of the (official) languages … and have an answer in the same language'. The wording of this article underlines the multilingual character of the Union. In their contacts with the European institutions and bodies citizens are free to use any of the official languages – not just their mother tongue.

Of course, European integration is not only or not even primarily about political interaction. The creation of a Union without internal borders, whose citizens have the right to live and work anywhere within the territory of the member states, can only succeed if people, institutions, and organisations can communicate, interact and co-operate economically and socially across language barriers and if people from different linguistic and cultural backgrounds view each other in a spirit of tolerance and mutual

understanding. The challenge inherent in this is tremendous. On the one hand, every effort must be made to allow people to retain their linguistic and cultural specificities, so that they will not see the process as a threat to their linguistic and cultural identity. On the other hand, European integration means growing trans-European co-operation and interaction across language and cultural barriers in all spheres of activity. Clearly, this has linguistic implications. Trans-European interaction can only work if people at large know other languages and if translation and interpreting are guaranteed at all levels. This is the background to the Union's endeavour over the past ten years to formulate a language-education policy, to promote language learning through concrete action – particularly through its action programmes in the fields of education and training – and to urge the member states to give greater priority to language teaching and learning in their educational systems. It is also the background to the Union's support for initiatives designed to improve the training of translators and interpreters.

The Union's ideas regarding language education are based on the conviction, first formulated in the White Paper on Teaching and Learning, 'that it is becoming necessary for everyone, irrespective of training and education routes chosen, to be able to acquire and keep up their ability to communicate in at least two Community languages in addition to their mother tongue' (p. 47) – the 1+>2 formula, in other words. Whereas the Council of Europe makes a terminological distinction between the multilingual character of Europe and the plurilingual competence of the individual, the Union favours the terms 'linguistic diversity' – for referring to European society – and 'multilingualism', for referring to a person's proficiency in a number of languages. Multilingualism, i.e. proficiency in at least three languages, is regarded as an important aspect of Union citizenship: 'Multilingualism is part and parcel of … European identity/citizenship' (p. 47). In the framework of this paper, it is impossible to describe the recommendations and actions of the Union in any detail. I should just like to add three points. (i) Through its action programmes, the Union seeks to give special emphasis to the less widely used and less taught among the Union's official languages. (It now also provides support for the learning of the languages of the candidate countries.) (ii) The European institutions are currently making renewed efforts to increase the mobility of young people, schoolchildren, students, researchers: all those being educated and their teachers. In this context, the development of multilingual skills is given specific emphasis – both in terms of the need for linguistic preparation and support for mobility and of the benefits to be derived from mobility. (iii) Language learning is regarded as a lifelong activity, extending from pre-school education to adult life.

To sum up, then, the following points can be made:

- the European Union is being constructed as a multilingual society; the eleven official languages of the Union are regarded as being equal in value

- language policy at EU level is aimed at promoting linguistic diversity and proficiency in a number of languages
- the notion of a *lingua franca* for the Union is rejected
- multilingual competence is an important aspect of EU citizenship; all citizens should have proficiency in a minimum of two Community languages in addition to their mother tongue
- multilingual competence is relevant for economic, political, and cultural reasons and is important for the development of mutual understanding among Europeans
- language teaching and learning are directly related to the political aim of European integration and constitute an important element of EU education policy; special support is provided for the less widely used and less taught languages
- translation and interpreting have an important role in maintaining and promoting linguistic diversity and are indispensable for getting business done in the EU
- there are no indications that language policy at Community level will change when the number of national languages increases yet again as a result of the expansion of the Union.

Regional and minority languages

From what I have dealt with so far, it will have become clear that language policy at EU level has little to say about the regional and minority languages of the European Union, such as Catalan, the immigrant languages that can be found in member states of the Union, such as Arabic, European languages outside the Union, the European Economic Area and the candidate countries, such as Russian, and about non-European languages that are relevant to the Union's dialogue and co-operation with other parts of the world, such as Japanese and Chinese. The explanation can be found in the limited remit of the EU in regard to language education: Article 149 of the Treaty of Amsterdam, which deals with 'Education, vocational training and youth' puts it like this. 'The Community shall contribute to the development of quality education by … supporting and supplementing (the member states') action … Community action shall be aimed at developing the European dimension in education, particularly through the teaching and dissemination **of the languages of the member states**.' This is generally interpreted as referring to the official languages of the member states.

The European Language Council

The Union's much less fervent commitment to the promotion of the regional and minority languages of the European Union has repeatedly come in for criticism. Article 22 of the Charter of Fundamental Rights of the European

Union, released last year, consists of one sentence: 'The Union shall respect cultural, religious and linguistic diversity.' There is a section on 'Regional and minority languages of the European Union' on the Commission's 'Education' website, which, among other things, lists projects that have received Community support. More recently, there have been indications of a movement towards including regional and minority languages in EU language policy and European Union programmes. The Opinion of the Committee of the Regions of 13 June 2001 on the Promotion and Protection of Regional and Minority Languages certainly points in this direction.

Needless to say, the universities' teaching, provision, development and research in the area of languages cannot solely or even primarily be guided by the political considerations of the European Union. I am convinced, however, that the promotion of societal and individual multilingualism as reflected by the language policy and language education policy propagated and promoted by the EU is of crucial importance for the future of the Union and that education in general, and higher education in particular, have responsibilities and duties in this respect; higher education in particular, because of the universities' wide-ranging language-related activities in teaching, provision, development and research, which include the following:

- modern language degree programmes, area studies programmes, and programmes combining language study with the study of other disciplines
- teacher education
- the training of translators and interpreters
- the delivery of courses or of portions of courses in one or more than one other language
- language provision for non-specialists, including linguistic preparation and support for mobility
- provision for university staff and for people from the non-university environment
- development of materials for the above types of programme and for language learning and assessment in general
- research related to the issues of linguistic diversity and of language learning, teaching, mediation, and assessment.

What should become clear from this list is that language study is not just relevant to language and language-related degree programmes, but potentially pervades the whole of higher education – just as the language issue is relevant to the whole of society. In fact, I believe it makes sense to view the above types of activity as being interrelated, constituting the 'area of languages in higher education', as it were.

The core of the above typology originates from one of the pilot projects that were carried out as precursors of the SOCRATES-ERASMUS Thematic Networks: the *SIGMA Scientific Committee on Languages* (12/94 – 10/95).

The members of the Committee were to produce national reports on the status quo in higher education language studies, to identify new needs and to propose measures for improvement and innovation. At the very first meeting, the Committee decided to focus on the transmission of linguistic and cultural knowledge and skills and on language mediation. This approach reflects one of the central aims of thematic networks in general: to address the frequently observed disconnection of higher education programmes from changing needs in the social, professional, and economic environments. SIGMA led to the creation of the Conseil Européen pour les Langues/European Language Council in 1997 and to the launch of the first fully fledged Thematic Network Project in the Area of Languages (1996–1999).

The CEL/ELC is a permanent and independent association of European higher education institutions and associations specialising in languages. Currently, its membership stands at 170. Its principal aim is quantitative and qualitative improvement in knowledge of all the languages and cultures of the EU and of other languages and cultures. It seeks to pursue this aim through European co-operation. Apart from conducting workshops, organising conferences, and publishing an information bulletin, the association has so far been active in policy development and in initiating European curriculum and materials development projects and co-operation projects, such as thematic network projects. Among the policy papers prepared was a proposal for higher education programmes and provision in regional and minority languages and immigrant languages and, more recently, a comprehensive document on the development and implementation of university language policies. This paper is intended as a framework for recommendations and actions in education and research linked to developments in the professional, economic, socio-cultural, and political domains.

The CEL/ELC co-operates with the European Commission, notably with the Directorate General for Education and Culture and the translation and interpreting services, the European Parliament and the Council of Europe as well as with associations such as Cercles, the European Association for International Education, and the European University Association. It carried out a pilot project for the trialling of the Council of Europe's European Language Portfolio in higher education, and the CEL/ELC and Cercles are about to develop a common Portfolio for use in higher education.

The Thematic Network Projects

The first fully fledged Thematic Network Project in the Area of Languages was a rather complex undertaking. We set ourselves the aim of bringing about reorientation in the programmes and provision mentioned above. The idea was to urge universities in general, and faculties, departments, and other units specialising in languages in particular, to respond to the challenges posed by

the construction of a multilingual and multicultural European Union – a tall order indeed. The project was structured into nine sub-projects devoted to specific types of programme or provision, such as postgraduate language studies, teacher education, the training of translators and interpreters, and language studies for students of other disciplines, and to transversal issues such as the less widely used and less taught languages, intercultural communication, new technologies, dictionaries, and testing. Some 130 experts from universities in the member states, Iceland, Norway, the Czech Republic, Hungary, Romania and Switzerland and from European associations, sitting on nine sub-project scientific committees, identified strengths and weaknesses in current policies, programmes and modes of delivery, developed recommendations (including course outlines and examples of good practice), and prepared or launched curriculum development projects, notably projects for the joint development of professionally orientated postgraduate modules and European Master's-type programmes. Among the projects launched from within the TNP were:

- a project for the joint development of a European Master's in Conference Interpreting
- a project for the development of an advanced level programme in multilingual education designed to prepare teachers for content-based language education at school
- projects for the development of continuing education modules for the use of information and communication technology in language study
- DIALANG – which is discussed elsewhere in this volume (see Chapter 8).

In addition, two books were prepared – on ICT and languages, and on intercultural communication.

The results of TNP1 were exploited and synthesised in a one-year dissemination project (1999–2000), which, for the first time, involved experts from all the countries in Central and Eastern Europe participating in the Socrates Programme and from Cyprus as well as representatives of large institutional associations. The results of TNP1 were regrouped under two broad themes and six sub-themes, as follows:

1. Language, mobility, citizenship
   a. European citizenship
   b. Mobility and co-operation: the needs of students and of the labour market
   c. New learning environments
2. Language Studies for Professional Life
   a. Training in translation and interpreting
   b. Education of teachers and trainers for a multilingual Europe
   c. Language studies at advanced level for non language professionals.

Six task forces prepared papers comprised of brief overviews of the respective issues, recommendations and examples of good practice. Together with an executive report, the papers are to be published later this year under the title 'Language Studies in Higher Education in Europe'. The document is to serve as a framework of reference for policy-makers and decision-makers across Europe.

Last September, we started a second three-year Thematic Network Project. It has three sub-projects devoted to Curriculum innovation – especially in first-degree programmes and continuing education, New learning environments and Quality enhancement. Each sub-project is to relate its work to three transversal issues: universities as actors in lifelong learning; the relevance of language studies to professional life; the European dimension. Again, each sub-project has a scientific committee made up of experts from universities across Europe. In Year One, committee members have prepared national reports, which will be synthesised. On the basis of the synthesis reports, initial sets of recommendations will be drawn up. The involvement of a large number of experts from Central and Eastern Europe has resulted in rather diverse findings. In Year Two, the committees will seek to address a wide range of academic and non-academic institutions and organisations outside the present project partnership, ranging from student organisations to the social partners, from academic and professional associations to language-testing consortia and organisations promoting language learning, mobility, and co-operation in Europe. The idea is to elicit comments on the recommendations and to undertake needs analyses. The recommendations, comments, and the results of the needs analyses will be presented for discussion at workshops, to which representatives of the various target groups will be invited. Year Three will see the preparation and launch of joint European projects related to the three issues covered by the project.

One might well ask what the impact of the TNPs has been to date. In some cases, there have been marked improvements. For example, the Master's in Conference Interpreting project launched from within the TNP has resulted in curriculum innovation across Europe and in the founding of a consortium of universities, who are determined to make innovation in this area an on-going process. Admittedly, the training of interpreters involves a relatively small number of institutions, students, and staff; moreover, the employers themselves told the universities clearly and loudly what qualifications were needed. It will be much more difficult to achieve similar reorientation in language teacher education, for example. However, one should not underestimate the reorientation in the minds of many of the experts involved in the TNPs; nor should one not underestimate the transfer of innovative ideas between universities and participating countries, which has led to the revision of programmes and the development of new programmes and new modes of delivery in many of the participating universities. Two difficulties remain:

how to reach out to higher education policy-makers and decision-makers and how to establish a dialogue with representatives of the social, professional, and economic environments. It is hoped that these difficulties can be at least partly overcome through co-operation with a multi-disciplinary TNP-type pilot project launched in May 2002.

There are two developments that may well speed up the process of innovation and improvement in higher education language studies: (i) the Bologna Process, and (ii) the dissemination of the *Common European Framework of Reference for Languages* and the large-scale introduction of the *European Language Portfolio* by educational authorities and institutions. The European Year of Languages 2001 could give added weight to these developments.

The Bologna Process is designed to make the universities themselves actors in the process of European integration. It envisages the creation of a European higher education area by 2010. It puts particular emphasis on the mobility of staff and students, on the competitiveness of European universities on the international stage, and on the employability of graduates. All three goals have linguistic implications. The Draft Berlin Declaration, which was issued by the Scientific Committee that prepared the Berlin EYL 2001 Conference, calls upon those carrying forward the Bologna Process to co-operate in developing multilingual and intercultural knowledge and skills as a precondition for the creation of a European higher education area. The Declaration makes quite specific recommendations, which combine EU policies, work undertaken by the Council of Europe, the Bologna Process, and ideas developed within the CEL/ELC and the Thematic Networks.

The Declaration expresses the view that multilingual competence, high levels of communicative competence in a number of languages, and language learning ability are becoming essential aspects of a European graduate's employability, citizenship, and personal development. It urges universities to provide students, regardless of their field of specialisation, with opportunities for improving their knowledge in languages, for learning new languages, and for becoming more independent in their language learning, thereby preparing them for lifelong language learning. To this end, universities must offer all students in undergraduate education the opportunity to take a number of credits in languages; they should create environments for independent language learning (including e-learning), encourage co-operative learning involving speakers of different languages, offer a wider range of languages (including the less widely used and less taught languages), and offer degree programmes or portions of programmes in other languages. As regards linguistic preparation and support for mobility, universities are reminded that they have a particular responsibility for promoting their own languages and for enabling incoming students to learn these languages to as high a level as possible. Language competence acquired through mobility should be assessed and certified.

Quantity, of course, is not everything. As becomes clear from the initial findings of the TNP2 sub-project on Quality enhancement, the issue of quality (one of the main concerns of the Bologna Declaration) has so far not been given proper consideration by those charged with language education at tertiary level. (The United Kingdom is an exception in this respect.) Language learning and teaching have to have clearly defined objectives related to language use. Objectives should be based on the Council of Europe's *Framework*. All language courses and modules should include assessment and have a credit value. This is to ensure transparency and comparability between institutions and languages and to facilitate credit transfer. It will enhance learner motivation and provide meaningful information for non-academic bodies such as future employers.

One of the objectives highlighted by the *Framework* and the *Portfolio* – and by DIALANG – is the acquisition of multilingual skills, rather than of unconnected proficiency in different languages: something that has yet to be grasped by those responsible for higher-education language studies. Needless to say, the development and delivery of this kind of language education requires highly qualified language specialists. Language education at university level has to be research-based.

As part of the Bologna Process, universities across Europe are engaged in introducing the bachelor-master system. Linked to the introduction of this two-tier system is the idea that first-degree programmes should equip graduates with the knowledge and skills that are essential for employment in the European labour market. In other words, Bologna is not just about new structures; it is also, and even more so, about course content. Regarding specialist language and language-related programmes, this means, among other things, that graduates have to have a high level of communication skills in the language or languages studied, and that they need to have acquired skills enabling them to adapt and further develop their linguistic repertoire in response to changing needs. This also means that programmes need to have a wider international dimension. For example, language majors will in future have to be encouraged, if not required, to study another European language alongside their major language.

Conclusion

During and after the Berlin Conference I gave a number of interviews to Germany-based newspapers and radio stations. Within no time, the interviews invariably turned to the question of language teaching and learning at school. According to EUROSTAT, only 53% of all adults in the EU can take part in a conversation in a language other than their mother tongue. Of course, we all know how this can be remedied: through early language teaching and through bilingual or multilingual education. This has become accepted wisdom. The

authorities at national or regional level in charge of school education have to take the necessary curricular measures and the universities have to provide the necessary programmes for initial and continuing teacher education to prepare teacher trainees and practising teachers for the new demands. In other words, major changes are required both at school and at higher-education level. I can hear you say, 'Yes, but this is going to cost money.' Right – but it is an investment in the future, Europe's future. We only have to look at the past to realise how necessary this investment is. Not only that. By enabling and encouraging children to learn languages we shall be fighting exclusion. In future, Europeans will only be considered literate if they are multiliterate.

I am labouring this point because I am aware that many people in Europe are beginning to question the validity of the goal of multilingual proficiency. For example, the European Parliament recently suggested that, in view of the EUROSTAT figures, one should focus on ensuring that 'every school-leaver is proficient in at least one European language other than his mother tongue'. The Director of the British Council in Germany argued at the opening of our EYL 2001 Conference that Europe would be well advised to opt for English as a *lingua franca*. As regards higher education, one can see that the triple process of EU expansion, the creation of a European area of higher education, and increasing competition on a global scale will strengthen the trend towards English as an academic *lingua franca*. In a sense, this is inevitable. It would also be wrong to ignore the fact that a high level of competence in English will be – already *is* – an important aspect of a European graduate's employability. However, as real and virtual mobility increases, more and more European citizens will be required to communicate in other languages. I for one cannot imagine that a doctor from the UK who has moved to Spain will get away with talking to his patients in English.

What all this means for the universities is that they have to consider their linguistic options carefully. Universities have to develop and implement their own specific and coherent language policies, covering the fields of education, research, and development. These need to reflect the European dimension, the specific needs of the non-academic environment, and institutional priorities and strengths. In doing so, they have to explore the added value to be gained from co-operation both within the individual institution and with other partners, nationally and internationally. And they have to pursue opportunities offered by co-operation with other sectors of education, cultural institutes, and other partners in public, private, and voluntary sectors.

Bibliography

Charter of fundamental rights of the European Union. *http://db.consilium.eu.int/df/default.asp?lang=en* *www.europarl.eu.int/charter/default_en.html*

Conseil Européen pour les Langues/European Language Council. Université et politique linguistique en Europe. Document de Référence. Berlin, 2001.

Conseil Européen pour les Langues/European Language Council. Website. *http://sprachlabor.fu-berlin.de/elc*

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council Resolution of 31 March 1995 on improving and diversifying language learning and teaching within the education systems of the European Union. *http://www.europa.eu.int/eur-lex/en/lif/dat/1995/en_395Y0812_01.html*

Council Resolution of 16 December 1997 on the early teaching of European Union Languages. *http://www.europa.eu.int/eur-lex/en/lif/dat/1998/en_398Y0103_01.html*

Decision No 1934/2000/EC of the European Parliament and of the Council of 17 July 2000 on the European Year of Languages 2001. *http://www.europa.eu.int/comm/education/languages/actions/decen.pdf*

The DIALANG Project website. *http://www.dialang.org/*

The Draft Berlin Declaration issued by the Scientific Committee of the Berlin EYL 2001 Conference (Freie Universität Berlin, 28–30 June 2001). *http://sprachlabor.fu-berlin.de/elc/docs/BDeclEN.pdf*

Eurobarometer Report 54: Europeans and Languages *http://europa.eu.int/comm/education/baroexe.pdf*

European Commission. Lingua website. *http://www.europa.eu.int/comm/education/languages/actions/lingua2.html*

European Commission. Memorandum on Higher Education in the European Community. Communication from the Commission to the Council on 5 November 1991 (COM (91) 349 final).

European Commission's website on regional and minority languages and cultures. *http://europa.eu.int/comm/education/langmin.html*

European Commission. 1995. White Paper on Education and Training – Teaching and Learning – Towards the Learning Society. *http://europa.eu.int/comm/education/lb-en.pdf* and *http://europa.eu.int/en/record/white/edu9511/*

Holdsworth, Paul. The work of the Language Policy Unit of the European Commission's Directorate-General for Education and Culture. CEL/ELC Information Bulletin 7 (2001): 11–15. *http://sprachlabor.fu-berlin.de/elc/bulletin/7/index.html*

Mackiewicz, Wolfgang. Six years of European co-operation projects in the area of languages. The Brussels Dissemination Conference. CEL/ELC Information Bulletin 6 (2000): 13–16. *http://sprachlabor.fu-berlin.de/elc/bulletin/6/index.html*

Mackiewicz, Wolfgang. Universities need a language lesson. The Guardian Weekly Learning English, June 21–27 2001: 1.

Mackiewicz, Wolfgang. Wie man seine Sprachkenntnisse im Internet testet. Fundiert 1 (2001): 90–97.

Opinion of the Committee of the Regions of 13 June 2001 on the Promotion and Protection of Regional and Minority Languages. *http://www.cor.eu.int/corz301.html* and *www.cor.ell.int/presse/PressRealeases2001/*

Resolution of the Council and the Representatives of the governments of the Member States meeting within the Council of 14 December 2000 concerning an action plan for mobility. *http://ue.eu.int/Newsroom/LoadDoc.cfm?MAX=1andDOC=!!!andBID=76 andDID=64245andGRP=3018andLANG=1* and *http://europa.eu.int/abc/doc/off/bull/en/200012/p000044.htm*

Sheils, Joseph. 2001. The Modern Languages Projects of the Council of Europe and the development of language policy. CEL/ELC Information Bulletin 7: 16–19. *http://userpage.fu-berlin.de/~elc/bulletin/7/index.html*

SIGMA Scientific Committee on Languages. Final Report. Berlin, 1995. *http://fu-berlin.de/elc/sigma/finalrep.html*

Thematic Network in the Area of Languages. Website. *http://fu-berlin.de/elc/en/tnplang.html*

Treaty Establishing the European Community. *http://europa.eu.int/eur-lex/en/treaties/dat/ec_cons_treaty_en.pdf*

Section 4

Work in Progress

11 TestDaF: Theoretical basis and empirical research

Rüdiger Grotjahn
Ruhr-Universität Bochum, Germany

Introduction

TestDaF (Test Deutsch als Fremdsprache – Test of German as a Foreign Language) is a new standardised language test for foreign students applying for entry to an institution of higher education in Germany. In this respect it is comparable to the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL). TestDaF consists of four sub-tests and takes 3 hours and 10 minutes to complete: 60 minutes for reading comprehension, 40 minutes for listening comprehension, 60 minutes for written expression, and 30 minutes for oral expression.

In the press, TestDaF has generally been compared to the TOEFL rather than to the IELTS. Characterising TestDaF as 'the German TOEFL' or 'TOEFL's little brother' is, however, misleading, seeing that in terms of content and format TestDaF is much closer to the IELTS than to the TOEFL.

One major reason for developing TestDaF was to make Germany more attractive internationally to foreign students who wish to study abroad by ensuring recognition of foreign applicants' language proficiency by German institutions of higher education, while the applicants are still in their home country.

To achieve this aim, the DAAD, the German Academic Exchange Service, set up a test development consortium[1] and financed the development of TestDaF from August 1998 to December 2000. During this period three parallel test sets were developed and piloted. In addition, various materials were developed, such as manuals for item writers and raters. Subsequently, the TestDaF Institute was set up at the University of Hagen to provide the necessary infrastructure for centralised test construction, centralised grading and centralised test evaluation. On April 26 2001, TestDaF's first international

1 The consortium consisted of the FernUniversität-Gesamthochschule in Hagen, the Goethe-Institut, Munich, the Carl Duisberg Centren, Cologne, and the Seminar für Sprachlehrforschung der Ruhr-Universität Bochum.

administration involved 20 countries. In 49 test centres licensed by the TestDaF Institute, more than 600 candidates took the test.[2]

At present, TestDaF is a paper-and-pencil test. In the medium term a computer-based version will be developed as well (for details, see Gutzat, Pauen and Voss 2000; Six 2000).

Description of the test format

TestDaF relates the candidate's performance in each sub-test to one of three levels of proficiency in the form of band descriptions termed TestDaF-Niveaustufen TDN 3, TDN 4 and TDN 5. The TDNs are intended to cover approximately the Council of Europe's Lower Vantage Level to Higher Effective Proficiency, or the bands 3 to 5 on the ALTE scale (see Association of Language Testers in Europe 1998; Council of Europe 2001; North 2000). This means that TestDaF is aimed at intermediate and advanced learners. No attempt is made to differentiate below TDN 3: it is only stated that the candidate has not yet reached the level of language proficiency required for admission to a German institution of higher education. In contrast to TOEFL, for example, no summary score for the candidate's overall performance is provided. (For a more detailed description of TestDaF see, for example, Grotjahn and Kleppin 2001; Gutzat, Pauen and Voss 2000; Projektgruppe TestDaF 2000; and *http://www.testdaf.de* .) The relationship between the Council of Europe scale, the ALTE scale and TestDaF's TDNs is shown in Figure 1 (c.f. TestDaF: Handreichungen für Testautoren 2000: 87).

Figure 1 scales and bands: Common European Framework of Reference, TestDaF, ALTE

Council of Europe (2001): Common European Framework of Reference:

A **Basic User**		**B** **Independent User**		**C** **Proficient User**	
A1 Breakthrough	A2 Waystage	B1 Threshold	B2 Vantage	C1 Effective Proficiency	C2 Mastery
		B1.1 B1.2	B2.1 B2.2	C1.1 C1.2	

2 At present, depending on where the applicant comes from, the fee for TestDaF varies from 90 to 110 Euros.

TestDaF:

TDN 3 TDN 4 TDN 5

Association of Language Testers in Europe (ALTE) (1998):

ALTE Level 1	**ALTE** Level 2	**ALTE** Level 3	**ALTE** Level 4	**ALTE** Level 5

In the case of the writing and speaking sub-tests, three versions of the TDNs exist: a short, test-user-orientated version intended for the candidate and those interested in the candidate's level of language proficiency, and two much more detailed versions intended for the raters and the item writers (c.f. Alderson's (1991) distinction between user-orientated, assessor-orientated and constructor-orientated scales). A translation of the band descriptions for Reading Comprehension reads as follows:

TDN 5

Can understand written texts from everyday university life (e.g. information on study organisation) as well as texts on academic subjects not related to specific disciplines (e.g. general environmental problems, socio-political issues), which are complex in terms of both language and content, with regard to overall meaning and specific details, and can also extract implicit information.

TDN 4

Can understand written texts from everyday university life (e.g. information on study organisation) as well as texts on academic subjects not related to specific disciplines (e.g. general environmental problems, socio-political issues), which are structurally similar to everyday usage, with regard to overall meaning and specific details.

TDN 3

Can understand written texts from everyday university life (e.g. information on study organisation) with regard to overall meaning and specific details; however, cannot adequately understand texts on academic subjects not related to specific disciplines (e.g. general environmental problems, socio-political issues).

In TDN 4 the same discourse domains are referred to as in TDN 5, namely everyday university life and academic subjects not related to specific disciplines (for a discussion as to whether to include subject matter knowledge as part of the construct to be measured, see for example, Alderson 2000, Chapter 4; Davies 2001; Douglas 2000). However, the texts are no longer characterised as complex in terms of both language and content, but as structurally similar to everyday usage. In addition, no extraction of implicit information is required. Finally, in TDN 3, complexity of information processing is further reduced by restricting the discourse domain to everyday university life.

With regard to admission to institutions of higher education in Germany, the following official regulations apply:

- Candidates who have achieved TDN 5 in all four sub-tests fulfil the linguistic requirements for admission to any German university
- In the case of candidates who have achieved TDN 5 in two sub-tests and TDN 4 in the other two, admission is possible in accordance with university-specific regulations
- Candidates who have achieved TDN 4 in at least two sub-tests and TDN 3 in the remaining sub-test(s) are admitted if in addition to their subject-matter studies they enrol on a German language course
- Candidates with TDN 3 or lower in three or four sub-tests have not yet achieved the necessary linguistic proficiency for studying in Germany and will not be admitted.

Exceptions to these regulations are possible in fields of study where the language of instruction is not German.

Reading Comprehension

The aim of the reading comprehension sub-test is to assess the candidate's ability to understand written texts relevant in academic contexts. The sub-test is intended to tap the following aspects of information processing: extraction of selective information, reading for the gist of a message and for detailed information, and complex information processing including comprehension of implicit information.

Three different types of text and task are used to test these skills: in the first task, a list of statements and several short and not very complex descriptive texts are to be matched. In the second task, a longer and more complex text is used, together with multiple-choice questions offering three options. The third task consists of a fairly long and complex text and forced-choice items of the type 'Yes/No/No relevant information in text'. Because the difficulty of a comprehension task is a function of the difficulty of both text and items, the complexity of both text and items has been taken into account in task design (c.f. Alderson 2000; Grotjahn 2000, 2001). There are 30 items in all. The level of difficulty of the three tasks is intended to match TDN 3, TDN 4 and TDN 5 respectively.

Listening Comprehension

The sub-test of Listening Comprehension is intended to assess the candidate's ability to understand oral texts relevant in academic contexts. Different levels of processing are tapped on the basis of three different types of text and item: selective information extraction, listening for gist and detailed information, and complex information processing.

In the first part, a dialogue typical of everyday student life at university is

presented once: the candidates are instructed to read the questions given beforehand and then to write short answers while listening to the text. In the second part, an interview or a discussion about study-related matters is presented once: the candidates are asked first to read the questions, which are in the true-false format, and then to answer them while listening to the text. In the third task, a monologue or a text containing relatively long monologue passages is played twice and the candidates have to write short answers to questions on the text. There are 25 questions in all. As in the case of Reading Comprehension, the level of difficulty of the three tasks corresponds to TDN 3, TDN 4 and TDN 5 respectively.

Written Expression

The aim of the sub-test of Written Expression is to assess the candidate's ability to write a coherent and well structured text on a given topic taken from an academic context. In particular, the following macro-skills are tested as both are key qualifications for academic study: (a) describing facts clearly and coherently; and (b) developing a complex argument in a well structured fashion.

In the first part of the writing task, a chart, table or diagram is provided along with a short introductory text, and the candidate is asked to describe the pertinent information. Specific points to be dealt with are stated in the rubric. In the second part, the candidate has to consider different positions on an aspect of the topic and write a well structured argument. The input consists of short statements, questions or quotes. As in the case of the description, aspects to be dealt with in the argumentation are stated in the rubric. Both parts have to be related to each other to form a coherent text.

The candidate's written text is assessed by two licensed raters on the basis of a detailed list of performance criteria. These include: grammatical and lexical correctness, range of grammar and lexis, degree of structure and coherence, and appropriateness of content. If the raters do not agree, the text is assessed by a third rater.

Oral Expression

The sub-test 'Oral Expression' assesses the candidate's ability to perform various conventional speech acts that are relevant in an academic context. The format of the sub-test is based on the Simulated Oral Proficiency Interview (SOPI) of the Center for Applied Linguistics in Washington, DC. Test delivery is controlled by a master audiotape and a test booklet, and the tasks are presented to the candidates both orally from tape and in print. The test is preferably done in a language laboratory or, if not possible, with two cassette recorders. The candidate's responses are recorded on a second tape, allowing centralised rating (c.f. Kenyon 2000).

The sub-test 'Oral Expression' consists of four parts and comprises tasks of varying levels of difficulty covering TDN 3 to TDN 5: in the first part, the 'warm-up', the candidate is asked to make a simple request. The second part, which consists of four tasks, focuses on speech acts relevant in everyday student life, such as obtaining and supplying information, making an urgent request and convincing someone of something. The third part, which consists of two tasks, centres on the speech act of 'describing', while the fourth part, which comprises three tasks, focuses on 'presenting arguments'.

The candidate's oral performance is assessed by two licensed raters on the basis of a detailed list of performance criteria. These include: fluency, clarity of speech, prosody and intonation, grammatical and lexical correctness, range of grammar and lexis, degree of structure and coherence, and appropriateness of content and register. If the raters do not agree, the performance is assessed by a third rater.

To reduce the number of tasks to be rated and thus make the rating process less time-consuming, the rating starts with tasks at the TDN 5 level. If the candidate's ability is considered to correspond to TDN 5, the rating process is terminated; otherwise, tasks at TDN 4 are assessed and it is decided whether the tasks at TDN 3 need to be rated as well (for a more detailed description of the sub-test 'Oral Expression' see Kniffka and Üstünsöz-Beurer 2001).

Test methodology

Each test set has been trialled several times: first, in an informal pilot study with approximately 20 German native speakers and approximately 40 learners of German as a foreign language in Germany; and second, with learners of German as a foreign language worldwide. In the latter case, the sample size varied from approximately 120 in the case of the writing and speaking sub-tests, to approximately 200 in the case of the reading and listening comprehension sub-tests.

In all trials the participants were given a questionnaire and asked to provide information on, for example, their language proficiency and periods of residence in Germany. In addition, the subjects could comment extensively on the test tasks. These data were intended to supplement the statistical item analysis by providing qualitative information on specific aspects of the test.

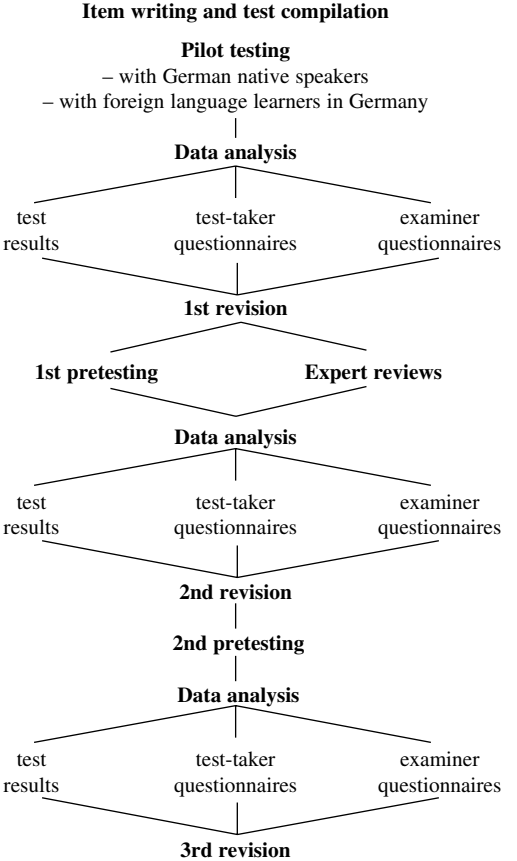
The examiners were also given a questionnaire in which they could comment on the administration of the test as well as on the test itself.

In the case of Reading and Listening Comprehension the pre-test results were statistically analysed by means of classical item analyses and Rasch analyses in co-operation with the University of Cambridge Local Examinations Syndicate (UCLES). Statistical criteria taken into account included: item difficulty, item discrimination, contribution of an item to sub-test reliability, item and person misfit. These criteria were applied flexibly in an iterative process.

In the cases of Oral and Written Expression the objectivity of the assessment was checked by analysing the inter-rater agreement by means of Cohen's weighted Kappa coefficient (c.f. Cohen 1968; Krauth 1995). In the future, with sufficiently large samples, the multi-faceted Rasch measurement model will be used to analyse specific aspects of the rating process such as rater severity and rater consistency (c.f. Embretson and Reise 2000; Linacre 1999; McNamara 1996, Chapter 5).

On the basis of the qualitative and quantitative information, items and texts were discarded or revised. Furthermore, various experts in the field of language testing and the teaching of German as a foreign language were asked to critically assess the test. A flow chart of the process of trialling is presented in Figure 2.

Figure 2 Process Model of Test Development (adapted from Gutzat, Pauen and Voss 2000)



The final versions of the three test sets turned out to have quite satisfactory psychometric properties. The reliabilities of the Reading Comprehension and Listening Comprehension sub-tests varied between approximately .70 and .85. As a rule, for high-stakes decisions on individuals a reliability of .9 or more is recommended. However, when the test results are reported in the form of a profile, as is the case with TestDaF, a reliability of .9 or more for each sub-test is normally not achievable, unless one considerably increases the number of tasks and items in each sub-test. This was not possible, because TestDaF is already quite lengthy. Nevertheless, this issue has to be considered in future revisions of TestDaF.

Another concern was the inter-rater agreement in the sub-tests 'Oral Expression' and 'Written Expression', which was in quite a number of instances not satisfactory, particularly in the case of 'Written Expression'. It is hoped that, with the help of the multi-faceted Rasch model, problematic aspects of the rating process could be better identified and the quality of the ratings thus improved.

Two other results of the statistical analyses also deserve mention. Both the classical item analyses and the Rasch analyses indicate that the first reading comprehension task, which requires a matching of statements and short texts, appears to tap a dimension separate from that measured by the two remaining tasks (for a justification see also Grotjahn 2000, p. 17f.; Taillefer 1996). As a consequence, scaling all items in the Reading Comprehension sub-test on a single dimension with the help of the one-parameter Rasch model appears to be problematical.

The other result worth noting relates to the yes/no questions in the Listening Comprehension test. Yes/no questions are quite common in listening tests. They have, however, some well known drawbacks. In addition to the high probability of guessing the correct answer, items with 'true' as the correct answer often discriminate best among low-proficiency learners, items with 'false' as the correct answer often differentiate best among high-proficiency learners, and neither of them discriminates well among low- and high-proficiency learners. Moreover, the correlation between items with 'true' as the correct answer and those with 'false' as the correct answer is often low or even negative (c.f. Grosse and Wright 1985; de Jong, December 24, 1999, personal communication). Because some of TestDaF's true/false-items behaved quite erratically, the true/false format should be reconsidered in future revisions.

Information technology

One of the roles of information technology in the TestDaF Project was to establish an appropriate infrastructure for test trialling and subsequent worldwide official administration. The information technology is conceived to support the current version of the test as well as the development of a

computer-based or even web-based format (for the latter, see for example Röver 2001).

The information technological infrastructure is being developed by the Department of Applied Computer Technology at the German Open University in Hagen (FernUniversität Gesamthochschule in Hagen), and consists of three main components, the item bank, the test-taker database and the test administration database. It can handle any type of data, including multimedia data (cf. Gutzat, Pauen and Voss 2000; Six 2000).

The **item bank** contains the items, tasks, sub-tests and tests as well as formatting instructions for the automatic assembly of tasks in a specific print format. An item or a task can be assigned certain attributes, which can then be used, for example, to search for specific items or tasks, to compile tests automatically or to select a task in computer-adaptive test administration.

The **test-taker database** contains personal information on the participants such as name, address, test date and level obtained in each of the four sub-tests. It can easily be searched and various statistical analyses can be carried out.

In the **test-administration database**, date and place of the examination are stored together with the candidate's scores and bands, the questionnaire answers and the results of item analyses such as difficulty and discrimination values.

For all three components a uniform, easy-to-use, web-based user interface is being developed on the basis of the most recent software technology to afford maximum flexibility as well as independence from commercial test creation and delivery software.

TestDaF's theoretical basis

In contrast to the TOEFL, for example, TestDaF is a criterion-referenced test: the candidate's scores are not interpreted with reference to the scores of the other candidates as in norm-referenced testing, but are related in all four sub-tests to the TestDaF bands 3, 4, and 5. While the sub-tests 'Written Expression' and 'Oral Expression' are direct, performance-based measures of writing and speaking skills, the sub-tests 'Reading Comprehension' and 'Listening Comprehension' are indirect measures of underlying constructs. As a consequence, criterion-referencing is quite straightforward in the cases of 'Written Expression' and 'Oral Expression', but rather problematical in those of 'Reading Comprehension' and 'Listening Comprehension'.

With regard to reading and listening comprehension the procedure is as follows: first, on the basis of the answers provided by the candidates, both person ability and item-difficulty are estimated with the help of the one-parameter Rasch model. Subsequently, each individual ability estimate is equated to one of the TDNs.

For the equation to the TDNs, grammar items from the German item bank of the computer-adaptive Linguaskill test were provided by UCLES as anchors.[3] The anchors had themselves been scaled and equated to the ALTE levels on the basis of scores from various examinations.

It is obvious that the characterisation of the candidate's reading and listening proficiency by means of a TDN is based on a highly complex chain of inferences: on the basis of the candidate's responses to the test items, a numerically based inference is made with regard to the non-observable construct 'comprehension ability in academic contexts'. Next, this ability estimate is related to a system of scaled can-do statements on the basis of quite limited data. Empirical studies are needed to demonstrate whether this highly inferential process as well as the descriptors used in the TDNs are sufficiently valid.[4]

When TestDaF and its theoretical basis were first presented to the public, a sometimes polemical discussion ensued. Political aspects aside, the following features of TestDaF were criticised in particular:

1. The testing of reading and listening as isolated skills rather than in combination with writing
2. The use of multiple-choice items rather than open-ended tasks
3. The testing of speaking by means of predetermined stimuli presented in print and on audiotape rather than by means of a face-to-face oral interview.

Before addressing each issue, I shall deal briefly with the question of authenticity in language testing, which is involved in all three issues.

Authenticity

It is often argued that language tests should be authentic – that is, that they should mirror as closely as possible the content and skills to be assessed. In my view, authenticity should not be overemphasised in the context of a high-stakes test such as TestDaF (for a similar view see Alderson 2000; Lewkowicz 2000). Authenticity might be important with regard to the face validity of a test or the potential impact on foreign language classes. However, a highly authentic test is not necessarily a highly valid test. If the candidate pays quite a large sum of money for a test and if success or failure in the test entails important consequences, the candidates will do their best even if they consider the test to be inauthentic.

---

3 In a recent study, Arras, Eckes and Grotjahn (2002) have investigated whether the C-Test could be used for the calibration of TestDaF's reading and listening comprehension items. The C-Test developed proved highly reliable (Cronbach's alpha = .84) and could be sucessfully scaled on the basis of Müller's (1999) Continuous Rating Scale Model. Furthermore, the C-Test correlated substantially with the reading, listening, writing and speaking parts of TestDaF (Spearman's $r > .64$).

4 Relating the content of the items clustering at a specific ability level to the TDNs proved to be not very informative (c.f. McNamara 1996, pp 200ff., for this kind of content referencing).

Testing of reading and listening in isolation

With regard to criticism that TestDaF assesses reading and listening comprehension as isolated skills, it can be argued that even if reading or listening entails an immediate written or spoken response, the actual process of comprehension takes place in isolation from writing or speaking. Furthermore, in academic contexts writing is often done quite a while after (extensive) reading has taken place. Moreover, listening, independent of speaking, is very critical for new students. For these and other reasons, it can be quite authentic to measure reading and listening independently of writing. Finally, the candidates themselves and other score users, such as admission officers or teachers in remedial German language classes, might be interested in information about specific strengths and weaknesses. However, when the measurement of reading and listening is tied to the measurement of productive skills, the single skills become blurred and can no longer be assessed validly.

Use of multiple-choice tasks

TestDaF's multiple-choice tasks have been criticised as being inauthentic and not able to tap higher comprehension processes. Yet several arguments in favour of multiple-choice items can be adduced (c.f. Bennett and Ward 1993; Haladyna 1999): (1) Multiple-choice tasks are necessarily more reliable than (extended) open-ended responses, because more items can be administered in the same period of time. (2) Seeing that more tasks can be administered, a broader range of content and skills can be covered. (3) The scoring of multiple-choice items tends to be more reliable than that of open-ended responses and is, in addition, much more economical. (4) Multiple-choice tasks lend themselves easily to computer-based and computer-adaptive testing. (5) Even highly complex processes (such as the comprehension of implicit information required in some of TestDaF's comprehension tasks) can be validly measured by means of multiple-choice questions. In view of these and other advantages, the inauthenticity of multiple-choice tasks appears to be a minor concern in the context of high-stakes testing.

Format of the sub-test 'Oral Expression'

It has been mentioned above that the sub-test 'Oral Expression' is based on the Simulated Oral Proficiency Interview developed at the Center for Applied Linguistics. The SOPI has now been in use for more than ten years and has been empirically investigated in quite a number of studies. It has, for example, been demonstrated that the SOPI correlates quite highly with the Oral Proficiency Interview (OPI) of the American Council on the Teaching of Foreign Languages (ACTFL), provided that the level of proficiency of the examinees is not too advanced (c.f. Kenyon and Tschirner 2000).

In the simulated oral interview, important aspects of genuine oral

communication are certainly missing and can thus not be adequately tested. However, as a consequence of the standardisation of the input and the centralised marking of the tapes, objectivity and reliability – and thus possibly also criterion-related validity – are higher than in a traditional oral proficiency interview. Furthermore, the SOPI is much more economical than the ACTFL OPI, at least if it can be administered as a group test in a language laboratory. Finally, as the recently developed computer-based version of the SOPI, the Computerised Oral Proficiency Instrument (COPI), demonstrates, with some slight modifications the SOPI format lends itself even to some form of computer-**adaptive** testing (c.f. Kenyon and Malabonga 2001; Kenyon, Malabonga and Carpenter 2001; Norris 2001).

Perspectives

In the case of a high-stakes admission test such as TestDaF, long-term, quantitatively and qualitatively orientated empirical research is a must (c.f. Grotjahn and Kleppin 2001: 429). A key issue to be investigated as soon as possible is the question of whether a candidate's TestDaF band profile is a sufficiently valid indicator of their communicative proficiency in a real-life academic context. Furthermore, research is needed for example into what a computer-based and a web-based version of TestDaF should look like. In this context, the following issues in particular need to be addressed in the near future:

1. Should the rather lengthy reading texts be replaced by texts that fit on a single screen? Or should the tasks in the reading and listening comprehension sub-tests even be replaced by a series of much shorter tasks, each consisting of a short text and a few items? Should these tasks then be treated as stochastically independent 'testlets' in the statistical analyses and be analysed by means of the partial credit model or Müller's (1999) Continuous Rating Scale Model? The 'testlet' approach would have the advantage that the reading and listening comprehension sub-tests could be made more reliable and that even a computer-adaptive test algorithm could be implemented. A possible drawback of such an approach is that the ability to comprehend complex and lengthy texts, which is important in an academic context, can probably not be adequately assessed in this way.
2. One should examine whether the marking of the sub-test 'Oral Expression' could be made less time-consuming. For example, one could investigate whether, in the case of a candidate identified as low-proficient in the reading and listening comprehension sub-tests, the rating should proceed in a bottom–up manner instead of the present top–down approach. With regard to a computer-based version, one should also investigate whether candidates identified as low-proficient beforehand should be given less difficult tasks than highly proficient candidates. Research into the COPI shows that oral proficiency testing can thus be made more economical and also less threatening for some candidates.

References

Alderson, J. C. 1991. Bands and scores. In J.C. Alderson and B. North (eds.), *Language Testing in the 1990s: The Communicative Legacy* (pp. 71–86). London: Macmillan.

Alderson, J. C. 2000. *Assessing Reading*. Cambridge: Cambridge University Press.

Arras, U., T. Eckes and R. Grotjahn. 2002. C-Tests im Rahmen des 'Test Deutsch als Fremdsprache' (TestDaF): Erste Forschungsergebnisse. In R. Grotjahn (ed.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (Vol. 4, pp. 175–209). Bochum: AKS-Verlag.

Association of Language Testers in Europe (ALTE) 1998. *ALTE handbook of European Language Examinations and Examination Systems*. Cambridge: University of Cambridge Local Examinations Syndicate.

Bennett, R. E. and W. C. Ward (eds.) 1993. *Construction vs. Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. Hillsdale, NJ: Erlbaum.

Chalhoub-Deville, M. (ed.) 1999. *Issues in Computer Adaptive Testing of Reading Proficiency*. Cambridge: Cambridge University Press.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70: 213–220.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

de Jong, J. H. A. L. (December 24, 1999). Personal communication.

Davies, A. 2001. The logic of testing Languages for Specific Purposes. *Language Testing* 18: 2, 133–147.

Douglas, D. 2000. *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

Embretson, S. E. and S. Reise. 2000. *Item Response Theory for Psychologists*. Hillsdale, NJ: Erlbaum.

Grosse, M. E. and B. D. Wright. 1985. Validity and reliability of true-false tests. *Educational and Psychological Measurement* 45: 1–14.

Grotjahn, R. 2000. Determinanten der Schwierigkeit von Leseverstehensaufgaben: Theoretische Grundlagen und Konsequenzen für die Entwicklung von TESTDAF. In S. Bolton (ed.), *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar* (pp. 7–55). Köln: VUB Gilde.

Grotjahn, R. 2001. Determinants of the difficulty of foreign language reading and listening comprehension tasks: Predicting task difficulty in language tests. In H. Pürschel and U. Raatz (eds.), *Tests and translation: Papers in memory of Christine Klein-Braley* (pp. 79–101). Bochum: AKS-Verlag.

Grotjahn, R. and K. Kleppin. 2001. TestDaF: Stand der Entwicklung und einige Perspektiven für Forschung und Praxis. In K. Aguado and C. Riemer (eds.), *Wege und Ziele: Zur Theorie, Empirie und Praxis des Deutschen als Fremdsprache (und anderer Fremdsprachen). Festschrift für Gert Henrici zum 60*. Geburtstag (pp. 419–433). Baltmannsweiler: Schneider Verlag Hohengehren.

Gutzat, B., P. Pauen and J. Voss. 2000. *Computer- und Interneteinsatz bei TestDaF*. Paper presented at the 21. Jahrestagung des Arbeitskreises der Sprachenzentren, Sprachlehrinstitute und Fremdspracheninstitute, March 10, 2000, Saarbrücken [available at: http://www.testdaf.de].

Haladyna, T. M. 1999. *Developing and Validating Multiple-Choice Test Items*. Mahwah, NJ: Erlbaum.

Kenyon, D. M. 2000. Tape-mediated oral proficiency testing: Considerations in developing Simulated Oral Proficiency Interviews (SOPIs). In S. Bolton (ed.), *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar* (pp. 87–106). Köln: VUB Gilde.

Kenyon, D. M. and V. Malabonga. 2001. Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning and Technology* (http://llt.msu.edu/vol5num2/), 5: 2, 60–83.

Kenyon, D. M., V. Malabonga and H. Carpenter. 2001, February 20–24. *Effects of examinee control on examinee attitudes and performance on computerized oral proficiency test*. Paper presented at the 23rd Annual Language Testing Research Colloquium, St. Louis MO.

Kenyon, D. M. and E. Tschirner. 2000. The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *The Modern Language Journal*, 84: 1, 85–101.

Kniffka, G. and D. Üstünsöz-Beurer. 2001. TestDaF: Mündlicher Ausdruck. Zur Entwicklung eines kassettengesteuerten Testformats. *Fremdsprachen Lehren und Lernen* 30: 127–149.

Krauth, J. 1995. *Testkonstruktion und Testtheorie*. Weinheim: Psychologische Verlags Union.

Lee, Y. -W. 2000. Identifying suspect item bundles for the detection of differential bundle functioning in an EFL reading comprehension test. In A. J. Kunnan (ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 105–127). Cambridge: Cambridge University Press.

Lewkowicz, J. A. 2000. Authenticity in language testing: Some outstanding questions. *Language Testing* 17: 1, 43–64.

Linacre, J. M. 1999. Measurement of judgments. In G. N. Masters and J. P. Keeves (eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 244–253). Amsterdam: Elsevier Science.

McNamara, T. F. 1996. *Measuring Second Language Performance*. London: Longman.

Müller, H. 1999. *Probabilistische Testmodelle für diskrete und kontinuierliche Ratingskalen: Einführung in die Item-Response-Theorie für abgestufte und kontinuierliche Items*. Bern: Huber.

Norris, J. M. 2001. Concerns with computerized adaptive oral proficiency assessment. *Language Learning and Technology* (http://llt.msu.edu/-vol5num2/), 5: 2, 99–105.

North, B. 2000. *The Development of a Common Framework Scale of Language Proficiency*. New York: Lang.

Projektgruppe TestDaF (2000). TestDaF: Konzeption, Stand der Entwicklung, Perspektiven. *Zeitschrift für Fremdsprachenforschung* 11: 1, 63–82.

Röver, C. 2001. Web-based language testing. *Language Learning and Technology* (http://llt.msu.edu/vol5num2/) 5: 2, 84–94.

Six, H. -W. 2000. Informatikunterstützung für TESTDAF. In S. Bolton (ed.), *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar* (pp. 131–140). Köln: VUB Gilde.

Taillefer, G. F. 1996. L2 reading ability: Further insight into the short-circuit hypothesis. *The Modern Language Journal* 80: 4, 461–477.

TestDaF: Handreichungen für Testautoren (2000). Bonn: DAAD.

12 A Progetto Lingue 2000 Impact Study, with special reference to language testing and certification

Roger Hawkey
Educational Consultant, UK

Introduction

An impact study of a major national language development innovation such as the Progetto Lingue 2000 (PL2000) in Italy, commissioned by the University of Cambridge Local Examinations Syndicate (UCLES), is an appropriate topic for inclusion in a collection of papers on language-testing issues in the European Year of Languages. This paper, which summarises and updates a presentation given at the ALTE Conference in Barcelona on 6 July 2001, outlines the first stages and presents some tentative early findings of the Cambridge PL2000 Impact Study. The presentation included video recordings of activities from Progetto Lingue classrooms, and of interviews with students, parents, teachers, school heads and Ministero della Pubblica Istruzione (MPI) officials. In addition, the PL2000 co-ordinator, Dr Raffaele Sanzo of the MPI (now renamed the Ministero dell'Istruzione e della Universitaria Ricera (MIUR)) described the principles and practices of the Progetto itself at the conference in a plenary presentation.

Impact and Washback

In the field of evaluation, impact is seen as 'the extent to which a programme has led to desired changes in the target field and/or audience' (McKay and Treffgarne 1998). Positive impact is considered to have taken place if learners are able to apply what has been learned with a beneficial effect and with relative ease. Impact may, of course, be planned or unplanned; positive or negative; achieved immediately or only after some time; and sustainable or unsustainable. Impact may be observed and measured during project implementation, at project completion, or some time after a project has ended (Kirkpatrick 1998; Varghese 1998).

In language teaching and testing, the concept of impact has been a matter of both theoretical and practical consideration, often in distinction from 'washback'. Hamp-Lyons (1997) sees washback as referring to the ways in which tests affect teaching and learning, and impact as covering their broader influence on education and society. Judged against these definitions, the Cambridge PL2000 Impact Study qualifies as an *impact* study, taking as it does the kind of multi-dimensional approach proposed by Bailey (1996), and Alderson and Wall (1993). The study considers the impact of the PL2000 on parents, educational managers, language-teaching materials producers, language testers and employers, and students and teachers. It also attempts to cover teaching/learning processes as well as content, what Milanovic and Saville (1996) refer to as the 'complex interactions between the factors which make up the teaching/learning context (including the individual learner, the teacher, the classroom environment, the choice and use of materials, etc.), …' (p. 2).

The Progetto Lingue 2000 Project

The Progetto Lingue 2000, the radical foreign-language improvement project with which the UCLES Impact Study is concerned, was preceded in the 1998/99 school year by the Ministry's *Linee Guida* project, which suggested that a foreign language should be taught outside the curriculum in lower secondary and primary schools. The favourable response to this project from teachers, students and parents led to the extension of the experiment to the whole of the national curriculum from the second year of kindergarten to the final year of secondary school, in the form of the PL2000 itself, initiated during the 1999/2000 school year under an agreement signed in January 2000.

The PL2000 aimed *to improve communication amongst Europeans and develop communicative competence in at least two languages of the European Union beside Italian (e.g. English, French, German, Spanish), making innovations in the teaching and learning of foreign languages in favour of the acquisition by students at every type and grade of school of verifiable, certifiable and cumulative levels of communicative competencies*. (Official MPI PL2000 booklet, 2000)

The Project has the following objectives:

- to modernise the teaching and learning of foreign languages
- to develop communication skills relevant to learners' language needs
- to encourage the achievement by learners of external certification at Council of Europe levels in two or more European languages
- to improve support systems such as resource centres and teachers' professional development.

Its main elements are:

- continuous foreign-language learning from kindergarten to the end of secondary school
- a second foreign language from year one of secondary school
- specified numbers of learning hours per level (e.g. 300 hours for the first foreign language at elementary- and middle-school levels; 240 hours for a second foreign language at middle-school level)
- teaching modules of 20–30 hrs with clearly stated objectives, contents and expected outcomes
- small homogeneous learning groups
- focus on communicative competence, especially listening and speaking
- the encouragement of self-access learning and the optimal use of information technology
- choice of internal or external examinations, with self-assessment and certified qualifications entered on individual European language portfolios to Level 2 on Council of Europe global scales
- provincial resource centres and extra teacher support, including PL2000-related in-service training.

UCLES Progetto Lingue Impact Study: background

The UCLES Progetto Lingue Impact Study (PLIS) was first proposed at meetings between Peter Hargreaves, Chief Executive of the University of Cambridge Local Examinations Syndicate (UCLES) English as a Foreign Language department, and Nick Saville, Head of the UCLES EFL Validation Unit, in September 2000. The PLIS, which was supported from the outset by Dr Raffaele Sanzo, the PL2000 co-ordinator at the Italian Ministero della Pubblica Istruzione (MPI), has the following aims:

- to ascertain the effects of the PL2000 language education reform project of the MPI on English-language performance targets, processes, media and certification
- to identify possible areas of PL2000 support
- to publicise the Italian model of adoption of the Common European Framework for language learning.

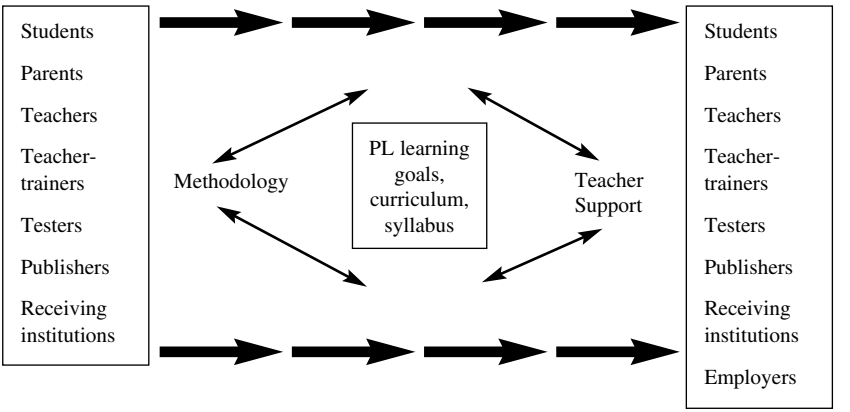
The PLIS, a longitudinal study to be carried out over two years, is co-ordinated by Roger Hawkey, external consultant to UCLES, working with Nick Saville and with Liam Vint, UCLES Development Manager in Italy. The study is supported by relevant UCLES departments on request, in particular Marketing, Video Unit, EFL Projects, Technical Support, and, in Italy, by the appropriate UCLES Local Secretaries.

The UCLES PL2000 Impact Study: approaches

The UCLES PL2000 Impact Study is investigating the impacts of the Progetto on a wide range of participants and other stakeholders, as mediated by programme curricula; methods, classroom activities, materials and media; tests and certification. The PLIS stakeholder and data-collection constituencies, that is elementary, middle and high schools, have recently been broadened by collaboration with the University of Siena Language Center, Centro Linguistico dell'Ateneo (CLA) and its *Idoneità Internazionale* (international qualification) project. This will ensure coverage of Progetto impacts on higher educational institutions receiving students from PL2000 schools.

Figure 1 suggests the main interdependent PL2000 components and the stakeholders on whom they may have direct or indirect impact.

Figure 1 PL2000 stakeholders and interdependent Project components



The Study benefits from research techniques developed by UCLES in its other national and international impact studies, for example the International English Language Testing System (IELTS) Impact Study (see Saville and Hawkey, forthcoming). Such studies have tended to support the view that impact is most effectively described by using a variety of data-collection techniques, both quantitative and qualitative, where possible achieving triangulation for the improved validation of findings. The PLIS is thus seeking both broad statistical and in-depth case-study data, the former through the UCLES data systems, the latter through the collection of qualitative data from a case-study sample of Italian schools at elementary-, middle- and high-school levels, in three regions (north, central and south) of Italy. Given some of the common goals shared by the PLIS and the IELTS Impact Study (IIS), also co-ordinated by Roger Hawkey, some of the validated modules and items from

IIS questionnaires inform the PLIS student and teacher data-collection instruments.

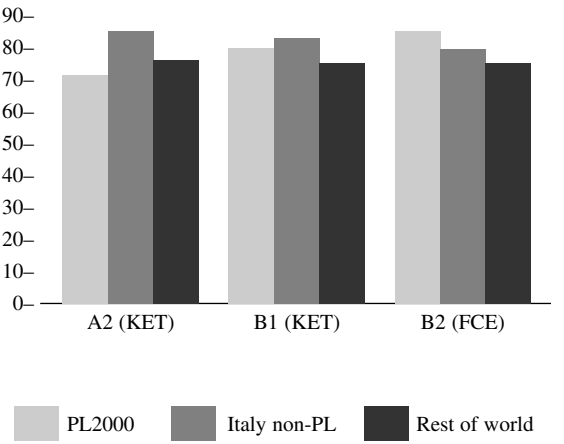
The PLIS has so far followed its original action schedule of preliminary trial visits to some of the schools in December 2000 and February 2001, followed by full-scale data-collection visits in October 2001 and April 2002. The analyses of these data will be included in a PLIS 2000-2002 *Report*, the findings of which will be a set of hypotheses about the PL2000, which may then be submitted for verification through broader samples of participants and other stakeholders.

The UCLES PL2000 Impact Study: data

The large-sample quantitative data collected and analysed for the PLIS will include statistics such as those already submitted to the MPI in the *Progetto Lingue 2000 Full Report: entry, performance and background data for the period January 2000 – July 2001*, by Liam Vint, UCLES Development Manager for Italy. All of these are potentially indicative of PL2000 *impact* because changes in candidate numbers and performance may well be found to be related to the Progetto. The data sets concerned include:

- external certification candidacies for UCLES examinations 1998–2001 (PL2000 vs total candidacies, by exam)
- PL2000 UCLES candidacies by Council of Europe levels (1998–2001)
- PL2000, Italy non-PL2000 and rest of the world pass-rate comparisons (see Figure 2 below, indicating a possible positive impact on exam results at Council of Europe level 2)

Figure 2 Pass-rate comparisons 2000/01
A2, B1 and B2 Level Exams



- PL2000 and other candidacies per school cycle (*media, licei classici, scientifici, magistrali; istituti tecnici, professionali, d'arte*)
- PL2000 and other candidacies per region
- students attending PL2000 exam preparation courses
- PL2000 and other external exam candidate ages, years of English study, hours per week of English, other foreign languages studied, English language exposure.

Relevant to the Impact Study, too, are the kind of written English language performance analyses for PL2000, Italy non-PL and global comparator students, which can be obtained from the Cambridge Learner Corpus (CLC) (see Appendix B). These have already been used in the pursuit of the teacher support aim of the Cambridge Impact Study (see above). Seminars using CLC data examine, with PL2000 teachers, selected writing exam scripts (suitably anonymised) of PL2000 candidates, with a view to improving future performance.

PLIS is also collecting and analysing longitudinal *case-study data*, including:

- video-recorded classroom observation of PL2000 teaching/learning approaches and activities, materials, media, classroom management and assessment
- semi-structured, video-recorded individual and group interviews with school heads, teachers, students, parents, alumni, PL2000 officials and employers
- the administration of the ALTE, Oxford University Press, UCLES Quick Placement Test (QPT) and the UCLES Language Learning Questionnaire (LLQ) to students in the case-study classes
- completions by students in the case-study classes of the PLIS student questionnaire on language background, foreign-language experience, use, attitudes and plans, testing experience, life goals
- case-study student work and internal test samples
- correspondence with PLIS teachers and students
- entries from teacher-contestants in the 2001–2002 UCLES PL2000 essay prize (Title: '*What the Progetto Lingue 2000 means to me as a teacher*').

By the end of April 2002, the longitudinal case-study visit data from the seven selected schools, an elementary school, a comprehensive school and a technical institute in the north, a middle school and a liceo in central Italy, and a middle school and science liceo in the south, included the following:

- 20 videoed PL2000 language classes
- 20 completed PLIS teacher questionnaires (see teacher questionnaire format at Appendix B)
- 110 completions each of the October 2001 and April 2002 PLIS student questionnaires (see student questionnaire format, along with school and

student profiles for one of the case-study schools, at Appendix C)
- 14 videoed interviews with school heads
- 20 videoed teacher interviews
- videoed interviews involving 20 parents.

More than 100 students are thus involved in the case-study classes, who will, in addition to their videoed classroom action, be covered by their own responses to:
- the PLIS student questionnaire
- the Quick Placement Test (QPT) to identify student proficiency levels corresponding to the Council of Europe Framework for Foreign Languages
- the Cambridge Language Learning Questionnaire (LLQ) on socio-psychological factors: attitudes; anxiety; motivation; effort, and strategic factors: cognitive and metacognitive

and by
- samples of their EL work
- internal language test performance data
- teacher comments.

UCLES PL2000 Impact Study: early indications

The PLIS still has important data-collection, corroboration and analysis phases to pass through. There are nevertheless interesting early indications, which are summarised here as examples of the kinds of impact that might be expected in the final analyses.

Our classroom videos show learner groups of various sizes, none exceptionally large, and some having relatively few hours per week of English. The means of grouping students for PL2000 language classes invite further examination. Some of our Quick Placement Test data raise questions on the language proficiency homogeneity of the classes, one of the objectives of the PL2000.

In its encouragement of school autonomy, the PL2000 does not, as many Ministry-led language improvement projects around the world have so often done, specify target curricular content. Rather, the PL2000 aligns its objectives with the target levels of the Common European Framework (CEF) (level A1 at exit from *scuola elementare*, A2 by the end of *scuola media*, and B1 after two years of secondary school). Qualitative data from the PLIS case-study schools so far suggest an interesting consequence of the intentional absence of a Progetto syllabus, and of the Progetto's encouragement of CEF target levels and external certification. This is the tendency for external *exams* to be regarded as the syllabus and the target criterion. It is commonplace to hear school heads, teachers and students refer to PL2000 courses by the name

of the target external exam (e.g. 'the PET course'). Such cases of impact from test to language course warrant in-depth analyses from our classroom, interview and questionnaire data.

PL2000 stands or falls, of course, on the effectiveness of the teaching and learning on PL programmes: how communicative *and how good* are the lessons? The Progetto certainly seems to have got its communicative message across although the communicative approach sometimes seems to be equated with oral skills training; this is perhaps a reaction against the reading and writing-biased language teaching of the past, and to the PL2000's implicit encouragement of appropriate *partial* competencies. Our analyses of the classes videoed are checking for coverage of *communicative* domains, purposes, settings, interaction, modes, media, levels, skills and functions. Video data so far suggest the following important areas of analysis in this regard:

- lesson planning and management
- task design and implementation
- variety of activity and interaction modes
- teacher : student relationships and activity balance
- information technology use and integration
- educational value added
- teaching/learning : test relationships.

There are mixed responses so far on the availability and use of the PL2000 resource centres; also on teacher support programmes. The financial support from MPI/MIUR for PL2000 courses is much appreciated, although there are sometimes problems of timing. The PL2000 agreement that students should not pay for PL2000-related exams, for example, may affect enrolments in external exams if fee financing is late. This has important implications for the end of the Project, too. Will fewer students enter for external exams when they or their schools have to pay?

There are clearly many more insights to come from the Cambridge PL2000 Impact Study. Even from the examples of early indications above, it may be seen that the areas of impact are broad and varied. Once the April 2002 follow-up data have been collected, there should also be evidence of *change* over a year or more of Progetto experience among our case-study students, parents, teachers, school heads and others.

References

Alderson, J. and D. Wall. 1993. Does washback exist? *Applied Linguistics* 14: 115–129.

Bailey, K. 1996. Working for washback: a review of the washback concept in language testing. *Language Testing* 13: 257–279.

Hamp-Lyons, L. 1997. Washback, impact and validity: Ethical concerns. *Language Testing* 14: 295–303.

Kirkpatrick, D. L. (ed.) 1998. *Another look at Evaluating Training Programs*. American Society for Training and Development (ASTD).

McKay, V. and C. Treffgarne. 1998. Evaluating Impact, *Proceedings of the Forum on Impact Studies* (24–25 September 1998), Department For International Development Educational Papers, Serial No. 35.

Milanovic, M. and N. Saville. 1996. 'Considering the impact of Cambridge EFL exams.'vv Research Notes 2, August 2000.

Saville, N. and R. Hawkey. (forthcoming). Investigating the washback of the International English Language Testing System on classroom materials. In Liying Cheng and Yoshinori J. Watanabe (eds). *Context and Method in Washback Research: The Influence of Language Testing on Teaching and Learning.*

Varghese, N. V. 1998. 'Evaluation vs. impact studies' in V. McKay and C. Treffgarne 1998. Evaluating Impact, *Proceedings of the Forum on Impact Studies* (24–25 September 1998), Department For International Development Educational Papers, Serial No. 35.

Appendix A

The Cambridge Learner Corpus

This 15 million+ corpus consists of Cambridge ESOL writing examination scripts.
The corpus is being built by Cambridge University Press and Cambridge ESOL for use by CUP authors, to inform their textbooks, and by Cambridge ESOL staff, mainly for test validation purposes.
Five million of the words in the corpus have been coded for learner errors.
The corpus is helpful for research into questions such as how learners use English, which areas of English cause the biggest problems for learners, how learners learn at different levels, and whether some errors are typical of particular language groups.
Contact: Fiona Barker, email: barker.f@ucles.org.uk
CLC can be found at http://uk.cambridge.org/elt/corpus/clc.htm

Appendix B

PLIS Case-study Teacher Questionnaire

About you

Your full name:

Form of address:	Miss	Mrs	Mr	Dr	Other
Your age, please:	20-30	31-40	41-50	51-60	61+

Institution where you work

Your position there

Your qualifications

Number of years teaching English

Types of school

About PL2000, your students and you

Please put a tick (√) in the boxes which show how often you implement the following student activities in your PL2000 English classes at school.	Frequently	Quite Often	Occasionally	Never	Comments
1. discussions in small groups					
2. pair discussions					
3. writing notes, letters or compositions					
4. doing grammar exercises					
5. discussions involving the whole class					
6. doing vocabulary exercises					
7. listening to the teacher talking to the whole class					
8. taking practice exams					
9. listening to recordings and taking notes					
10. reading then writing answers to questions					
11. reading texts (from books or other materials)					
12. listening and choosing answers to questions					
13. discussing the external exam					
14. + Other activities in your English classes:					

Please put a tick (√) in the boxes which show how well the following objectives of the PL2000 have been achieved in your school.	Very well	Well	Not very well	Hardly at all	Comments
1. modernising the teaching and learning of foreign languages					
2. language teaching through modules with clear objectives, numbers of hours, content, expected outcomes					
3. small homogeneous language-learner groups					
4. communication skills relevant to students' language needs					
5. more self-access language learning					
6. optimal use of modern technology in language learning					
7. the use of external language exams at recognised Council of Europe levels					
8. improved teacher support such as resource centres and teacher professional development					

Please list here and comment on what you see as the *advantages and disadvantages* of the PL2000 for your students and you.

Please write here how you think your English language teaching and attitudes have c*hanged over the past year*.

PLIS TQ
Thank you very much!

Appendix C

PLIS Case-study Student Questionnaire

About you and your language learning

Full name			
School	Class	Age	Male/Female
Language(s) you study at school?			
How many years have you studied English?			
Any *extra* English course(s) this school year? (Yes/No)			
If yes, what course(s)? Where, when?			

Please answer these questions about the use of English outside your school classes.	Yes/No	If yes, where? When and how long?
Have you ever studied English abroad?		
Have you used English as a tourist?		
Have you ever used English in a job?		
Which English tests or examinations have you taken?		
Which English tests or examinations will you take in the future?		
What job/career are you aiming at for your future?		
Do you intend to study at university? If so, what subject(s)?		

Please put a tick (√) in the boxes which show how often you to do the following activities in your English classes at school.	Frequently	Quite often	Sometimes	Never	Comments
1. discussions in small groups					
2. discussions with a partner					
3. writing notes, letters or compositions					
4. grammar exercises					
5. discussions with the whole class					
6. vocabulary exercises					
7. listening to the teacher talking to the whole class					
8. taking practice exams					
9. listening and taking notes					
10. reading then writing answers to questions					
11. reading texts (from books or other materials)					
12. listening and choosing answers to questions					
13. discussing the external exam					
14. + Other activities in your English classes:					

Please put a tick (√) in the column (a, b, c, or d) for the choice that best describes your response to each of the statements below.	a	b	c	d	Comments
1. My English has improved this school year a. A lot. b. Quite a lot. c. A little. d. Not at all.					
2. My English language skill that has improved most is a. reading. b. listening. c. writing. d. speaking.					
3. Most school days I spend …… studying English outside class. a. less than one hour b. one to two hours c. more than two hours d. no time					
4. My main feeling about taking *external* exams in English is they a. are more difficult than school exams. b. help me to work harder. c. are for listening and speaking. d. are necessary to enter university.					
5. My main reason for learning English is a. to communicate with more people. b. to get a better job. c. for travel overseas. d. because I like it.					

Please put numbers in the boxes to show how often you do the following *activities in English outside your school.*
0 = never; 1 = almost never; 2 = occasionally; 3 = often

reading books		reading newspapers, magazines		writing letters		using email		watching TV	
listening to the radio		using the Internet		watching movies		going to shows		talking with visitors	

Other activities using English? *please specify*

Please write here how you think your English language learning and attitudes have changed over the past year.

Thank you! PLIS SQ1A04/02

Sample school profile

School Profile (3)

Liceo Ginnasio, Virgilio, ROMA	N805	Cl=43	T=86 9/2	ELT/Ass= 15	PLICl= 9	PLE=	hpw= 1 am–1.5 pm

FL	E, F, G, SP
Tests	Cl 1 elementary; 2, pre-inter, 3 intermediate, 4 upper-inter, 5 mostly Brit Lit
TBS	Headway; Enterprise; FC Gold; Now and Then (Zanichelli; Views of Literature Loescher)
R Centre	Yes; 2nd floor: books, videos, cassettes, magazines; becoming TRC
Use of E	TV, Internet, overseas travel, study holidays; further studies, travel
PLI	+ co-operation, objectivity in evaluation competence, student motivation

Case-study Student: Class Ages

18	17	16	15	M/F?	Class
5	1	10	2	5/13	3F

About your languages

What language(s) do you usually speak at home?

Italian 18 Polish 1

What language(s) do you study at school?

English 18 French 2 German 0 Spanish 3

How long have you studied English?

5 years 1 6 yrs 2 7 yrs 7 8 yrs 4 10 yrs 4

How many hours of school English classes a week?

School hrs 3

Extra English learning? What? No

Which English tests or examinations have you taken?

PET 1

Trinity 1

Which English tests or examinations will you take?

FCE 8

Have you ever studied English abroad? Y/N

Yes 9 No 9

If so, where?

Ireland 5	2000,// 2001,	3 weeks////,
	2001, 98,99	2 wks, 2 wks
Malta	2001	2 wks
England 7	2001, 2000///,	1 week, 2 wks///
	98//, 99/	3 weeks //l/

Have you had other experience of English? Y/N

Yes 2 No 21

For example: European education projects, if so, please specify.

Yes 1

As a tourist? Y/N

Yes 4 No

Where?

Canada 2 Hungary 1 NL 1 France 2 Austria 1 Greece 1 US 4 England 5 Ireland 2

Working?

Yes 0 No 18

Jobs?

No 18

Self-access English

	0	**1**	**2**	**3**		**0**	**1**	**2**	**3**
books	5	7	5	1	newspapers	5	9	4	0
magazines	5	8	1	4	letters	7	3	4	5
email	8	0	6	4	Internet	4	4	2	8
TV	2	10	6	0	radio	9	6	3	0
cinema	6	9	3	0	talking with visitors	4	3	10	1
shows	14	1	1	2					

Others: pop songs 1 music 1

friends at school/family friends

What are your opinions about learning English?

For work 8	Holidays, travel 3/	Contacting others 3	World language 6
Studies 0	Liking 1	Culture 0	Difficult to teach 1
For the future 1			

What job/career are you aiming at for your future?

Cook	Archaelogist 1	Teacher	IT specialist
Actor 1	Tour operator	Air hostess/model	Engineer
Vet 1	Sports star	Lawyer	Lorry driver
Musician	Psychologist 1	Interpreter 1	Photographer 1
Journalist 2	Doctor 1	Biologist 1	Don't know 11

Do you intend to study at university?

Yes 16 No 2

If so, what subject(s)?

No specification 10	Dramatic arts 0	Biology 1	Journalism 1
Medicine 1 Anthropology 1	Languages 1	Archaeology 1	Law 1

13 Distance-learning Spanish courses: a follow-up and assessment system

Silvia María Olalde Vegas and Olga Juan Lázaro
Conclusions by Milagros Ortín Fernández-Tostado and Javier Fruns Giménez
Instituto Cervantes

Translation by Maureen Dolan

Introduction

The easy-access Assessment System of the Instituto Cervantes' distance-learning Spanish courses on the Internet allows students to consult a large, well-organised storage area housing all the information on their performance.

The system was designed for semi-autonomous use and to allow students to develop their own learning strategies, working jointly with study group members and tutors. It falls to students themselves to use the system as often as necessary to gain a clear picture of their progress.

The Follow-up and Assessment System is housed in the Study Room, the place reserved for students to work in with their group companions and tutor. Here students have all the teaching aids and materials at their fingertips. This space is both the students' reference and starting point.

To receive assessment information, students have only to click on the Assessment System icon to consult a number of reports organised into three systems or sections: Automatic Follow-up, Automatic Progress Test and Course Tutor's Assessment.

Automatic Follow-up

This system, activated by students themselves, accesses a wide-ranging document. Whenever students require an overall view of their performance and work pace or wish to revise certain points, the Automatic Follow-up System provides the requested information in a stratified and detailed manner.

At the behest of students, this system stores the results of the study exercises done by students on their own, without help from tutors or companions.

Automatic Progress Test

The Automatic Progress Test automatically collects and stores each student's performance in exercises we have dubbed 'Passport Controls'.

These progress-test pauses confront students with a variety of exercises summing up the contents of the level already studied. The function of the Passport Controls is to reassure students that they are on the right road towards acquiring the course skills.

Course Tutor's Assessment

Our distance-learning Spanish courses are designed not only to permit individual students to use the computer system, but also to facilitate communicative interaction between the individual student and members of his or her group, and also with the course tutor.

Communication among students within the same group, or occasionally between different groups, is achieved via a series of activities called Round-up Exercises and End of Lesson Tasks.

These exercises generate work to be assessed by tutors themselves. It is the tutor, rather than the computer system (as occurs in the cases outlined above), who assesses the communicative abilities displayed by the student when carrying out these activities.

The tutor's assessment can include suggestions for improving performance. For this reason, tutors should encourage students to frequent this site. Students can expect to receive not only the tutor's evaluation, but also his or her guidance and advice.

Theoretical Framework

In the wake of the Internet, a new medium which has generated new needs, the Instituto Cervantes has created its distance-learning Spanish courses on the Internet (Gerardo Arrarte *et al*. 2000). Our objective is to teach Spanish as a Second Language to an adult public by using this new technology.

We believe that our target students, as students of a second language, expect to be able to make satisfactory communicative exchanges in the target language, and, as students who have opted for a distance-learning method on the Internet, they expect to be able to acquire this new knowledge within a flexible timetable.

These two approaches, at first glance very different, interlock perfectly in the Instituto Cervantes' distance-learning Spanish courses.

For two reasons, our courses follow the task-based learning principles of the Communicative Approach. Firstly, the Internet permits the use of real materials and encourages students to seek out new examples of the target language. Secondly, contacting other students who share a similar interest in

the cultural and linguistic diversity of Spanish is relatively straightforward.

The distance-learning Spanish courses are structured into lessons, which are the main didactic unit. These lessons build towards an End of Lesson Task, in which the group of students must create a given product (a leaflet, a restaurant menu, a travel guide, a file containing fellow-students' details, etc.). The End of Lesson Tasks are preceded by a series of Round-up Exercises that function as gateway tasks.

When we talk about the Communicative Approach, we stress that students are expected to interact with their companions or, where this is not possible, with their tutor. We refer to students as companions because, when they matriculate, they join a collaborative study group. To a certain extent, group members share similar interests and communicative limitations, and are therefore co-ordinated by a course tutor.

This collaborative study group stems from the open, distance-learning approach in which these courses are rooted. As A. W. Bates explains (1995: 27):

> Open learning is primarily a goal, or an educational policy: the provision of learning in a flexible manner, built around the geographical, social and time constraints of individual learners...

He goes on to stress that:

> Distance education is one means to that end: it is one way by which students can learn in a flexible manner...

The students in a group share the social, geographical and time limitations mentioned above, as well as, in the case of foreign- or second-language acquisition, linguistic and socio-cultural limitations.

Professionals in the field are studying seriously the changes brought about by this new distance-learning mode and the impact on it of new information, training and idea-exchange tools such as the Internet. There is already an abundance of literature on new student and teacher or tutor roles; the new functions which must be taken on board by the participants in the teaching and learning process. However, wide-ranging experimentation or field studies are still lacking.

By 1979, H. Holec (1979: 25) was already describing the new role of the tutor in independent learning situations. Basically, the tutor's role is that of guide and counsellor because it is the student who sets the pace of study. Consequently, however, this also places responsibility for the student's success and that of his or her companions at the student's own door, given that each person's participation is necessary to bring the Round-up Exercises and End of Lesson Tasks to a successful close. Given this arrangement, the Spanish course student also overcomes the isolation that can be an obstacle in distance-learning.

This idea is expressed by A. Giovannini *et al*. (1996: 27) as follows:

> anyone who wants to learn a foreign language must be aware that the learning process depends to a large extent on their own sense of responsibility and degree of participation.

Despite this requirement to work together, students still benefit from the advantages of the one-to-one attention and assessment of a tutor, responsible for assessing communicative exercises. They also benefit from the tools offered by the computer system for the more quantifiable aspects.

The principal aim of our present analysis is to showcase the assessment tools we have incorporated into the computer system to speed students' acquisition of knowledge and different communication skills, and to show how these tools aid tutors in this new learning area.

Finally, returning to the elements we have tried to keep in mind in the design of the distance-learning Spanish courses on the Internet, we stress again the fact that the students have other group companions around them to offer support and a tutor leading the way.

Joint interaction is achieved through the use of the communication tools provided by the Internet. This communicative exchange can take place orally, in written form, through the use of communication tools such as email, chat, forums, audio and video-conferencing ... and whatever this fast-moving technology comes up with next. With this in mind, we have tried to create an environment capable of incorporating changes into this new medium, which is often difficult to keep up with.

Our distance-learning Spanish courses arise from a mutual adaptation process involving learning tasks, the aforementioned tools, the group of students and their tutor. This said, we must now consider the real significance of this communicative exchange in and of itself.

The technical aspects at work, making this exchange possible, allow us to embrace the notion of electronic literacy (C. A. Chapelle 2001: 88):

> As language learners are increasingly preparing for a life of interaction with computers and with other people through computers (D.E. Murray, 1995), their 'electronic literacy' (Warschauer, 1999) becomes an additional target. An argument about authenticity needs to address the question of the extent to which the CALL task affords the opportunity to use the target language in ways that learners will be called upon to do as language users, which today includes a variety of electronic communication.

As for the tutor, his or her role corresponds to the profile of guide and counsellor, as described by Holec, cited above (1979: 25). The tutor helps to draw out the linguistic and communicative elements which constitute the aim of study and encourages a satisfactory work pace. The level of commitment is paced by students themselves but it must also, in turn, keep up with the pace of the group. If they wish, then, Spanish course students can overcome the isolation which is often inherent in distance-learning.

Added to this is the fact that, in this kind of environment, the student has to

make the effort to participate. To paraphrase Giovannini *et al.*(1996: 27), anyone who wants to learn a foreign language must be aware that the learning process largely depends on his or her own sense of responsibility and degree of participation.

These are the principles that have inspired us to design a tailor-made Assessment System, described below, for these Spanish courses. Our aim is to offer students '...detailed guidance and contextual information on their learning process' (Soria y Luzón 2000). We trust that our efforts will serve as the basis of an ever-improving assessment model, capable of meeting the requirements of students of Spanish on the Internet.

Assessment System

Data storage

Students deal with our courses in an independent, autonomous manner and require tools informing them on how much has been completed and how well. At the outset of the course, students may remember what they have already done, but after completing a number of exercises over a number of hours, or if work has been held up for some time, students will need reminders. The Assessment System was created to meet this need.

The minimum storage unit built into the design of the course is the exercise (although an exercise can be made up of more than one computer screen page). At the end of each exercise, an icon comes up to allow storage of data and results.

It allows the system to pick up the student's participation in the exercise, evaluate and store the correct answers (further on we will find out what this is based on) and indicate the time spent.

If the students are not satisfied with their performance they can repeat the exercise. The computer also registers the number of attempts.

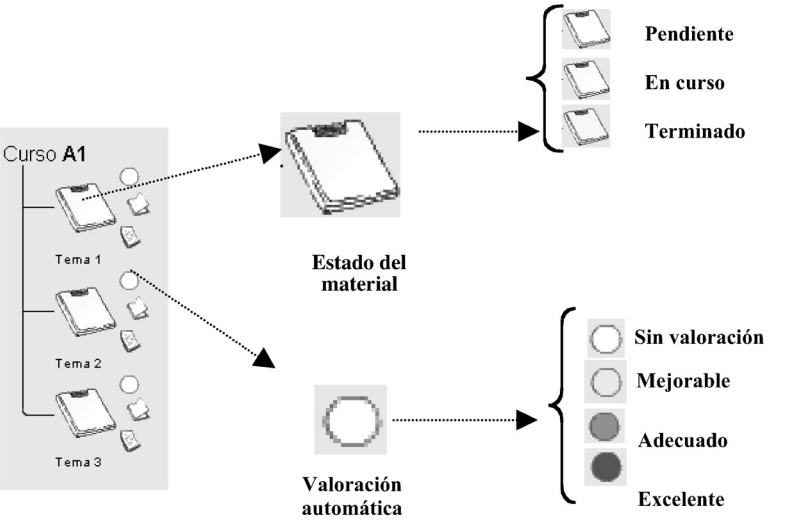
Therefore, it is the students who control the option of storing their work on a particular exercise, along with the results obtained. This solution was chosen over automatic storage of student participation to make students themselves responsible for their own learning process. Proper progress depends on an awareness of this fact.

By clicking on Assessment, an information table, which can be opened if required, appears on the screen. The information it contains is twofold, as can be seen below:

> The material covered, detailing what has not yet been accessed (waiting), what has been accessed but not completed (under way) and what material has been fully dealt with (completed).

> An automatic assessment of how the different exercises comprising each piece of material have been done. This information is made available in a simple colour code.

Figure 1



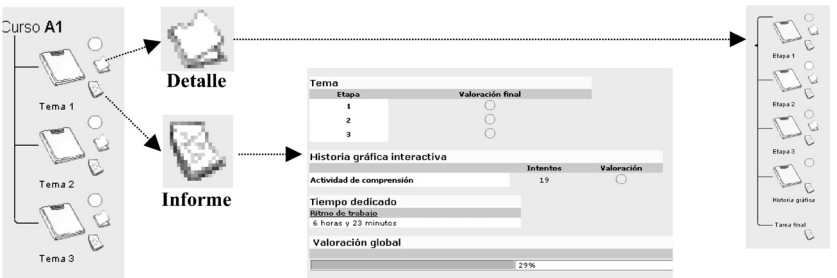
The information that the student receives in this figure deals with, on the one hand, the status of the material that has been worked on, indicating lesson by lesson whether the entire study route has been completed.

On the other hand, it presents the automatic assessment of what has been completed up to that point, consisting of the percentage of correct answers obtained in the stored exercises.

To this is added the opportunity of receiving more detailed information on the lesson or the option of going back down to the next information level corresponding to the stages of each lesson.

We can take a closer look.

Figure 2



By clicking on the 'Details' icon and then the 'Report' icon, the student can:

- Run down the table of contents showing the course structure to find the level he or she is interested in.
- Access the Assessment Report corresponding to the content of the selected level.

The *Report* opens a table showing the student's performance from different perspectives.

The *Final Evaluation* reflects the percentage of correct answers obtained in the exercises stored by the student in each stage.

The *Interactive Graphic Story* shows the number of attempts and the evaluation of this material, which revises the lesson contents in a game format.

The *Time Spent* allows the student to refer to the time spent on the courses and compare it with the marks obtained.

The *Final Evaluation* offers a compendium of the automatic follow-up marks stored by the computer for the student in each lesson.

Student commitment

When learning a language, communication is fundamental, as is, therefore, contact with other students or speakers.

We believe this to be just as important in distance-learning. It is fundamental to set up a virtual learning community – a group – and encourage the formation of motivating relationships among its members to help them achieve their study aims (we also learn from other people outside this group).

As J. Fernández notes (1999: 78):

> the imparting of knowledge also occurs from person to person, and so another of the great benefits of the Internet is this: interactivity. The world of education is ever more dynamic because now the students can spread knowledge in ways never before imagined.

In the words of C. A. Chapelle (2001:32):

> the experience crucial for individual cognitive development takes place through interaction with others, and therefore key evidence for the quality of a learning activity should be found in the discourse that occurs in the collaborative environment.

For this reason, we have created a tool that provides information on the student's work pace and that of his or her group companions. This can be accessed from the Automatic Follow-up option of the Assessment System.

When the student clicks on this tool, the computer provides illustrated information in two coloured lines. One line represents the work pace of the student requesting the report, whereas a second line in a different colour

represents the work pace of the study group to which he or she belongs.

In making this comparison possible, we have two aims in mind. On the one hand, the students are encouraged to establish a comfortable work pace, in the knowledge that they are the only people who can do this, in keeping with their own particular circumstances. In any case, this is also unavoidable in the kind of language course we have designed, because the communication exercises, which are carried out simultaneously (via chat, for example) or time-delayed (email), require answers that emerge from the didactic sequence, if progress is to be made in the acquisition and mastery of linguistic and other contents.

On the other hand, the tutor, as guide, can immediately spot any students who fall behind the group pace, and offer them support. As described in the previous paragraph, the tutor may even find himself with another function, namely that of work partner in the programmed communication exercises (if a student falls behind the rest of the group and has no one with whom to do the exercise).

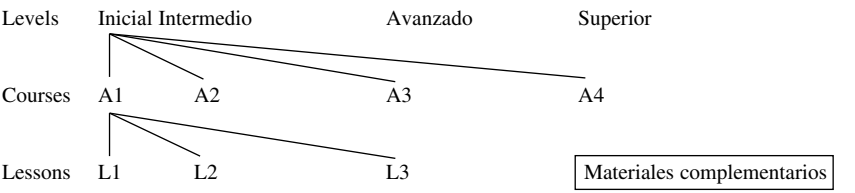
This tool is geared only towards informing the student and backing up the idea of group learning. At all times the student has a motivating point of reference in his companions, thus avoiding any sense of isolation.

Automatic Progress Test

The structure of the Spanish courses

The Spanish courses are structured into the four levels set by the Curricular Plan of the Instituto Cervantes. Each level is divided into four courses and these are further divided into lessons. See the diagram below.

Figure 4 The structure of the Spanish Courses



Each lesson is built around ten work sessions lasting roughly an hour each. The design of these work sessions includes prior presentation of new linguistic and socio-cultural content and practice exercises graded from the more controlled to the more open. At the end of each session, the student has to contact the rest of the group to employ what he or she has learned in real communicative interactions or exchanges.

Any student dealing alone with the acquisition of knowledge has to be sure he or she is working correctly and that acquisition is occurring effectively.

To meet this requirement, Passport Controls or progress tests have been built in every three work sessions. Therefore, there are three Passport Controls in each lesson.

Passport Control

The Passport Controls are designed as a sequence of interactive exercises of a variety of types (dragging, choosing, writing, etc.), in which students have to show the knowledge and skills they have acquired (reading comprehension, listening comprehension and writing skills). The more open oral communication and written expression skills are assessed by the tutor.

This test differs from the other work-session interactive exercises, automatically corrected by the computer, in that it provides a percentage mark once all the exercises have been completed. Right and wrong answers are not indicated. The underlying intention is to allow the students, in charge of their own learning process, to go back and revise the exercises in which they have obtained the poorest results (they can see this from the detailed report in Automatic Follow-up). This done, they can return to Passport Control and try to increase their percentage of correct answers.

In this approach we have taken as our lead authors such as A. Giovannini *et al*. (1996: 29):

> autonomy depends on our decision-making capacities and our ability to accept responsibility, to self-evaluate and supervise our own learning process...

In short, it is up to the students themselves to check whether they have understood what they have studied.

As with the storage of exercise work, the students can decide to do the Passport Control exercises without storing their answers if they are not ready to have their attempts registered by the system.

Evaluating Passport Control

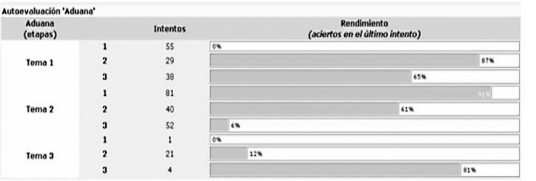
Depending on the results, the student will see a rating scale with an evaluation depending on the percentage of correct answers obtained: improvable, satisfactory and excellent, visually represented in a green band of increasing intensity. The student also receives an automatic recommendation from the computer, which basically either advises him or her to revise the contents of the relevant sessions or to continue his or her study sequence by going on to the next exercise.

If a student wishes to have an overall view of his or her results in the different progress tests, he or she must access the Follow-up System in the Study Room and choose Self-correction, where the nine course Passport

Controls are shown.

The computer informs the student of the number of times he or she has done each Passport Control, in the column marked 'Attempts', and of the percentage of correct answers obtained on the latest try, as can be seen below:

Figure 5



Course Tutor's Assessment

The Round-Up Exercises and the End of Lesson Tasks

Evaluation of learning is a key element of the teaching and learning process of the Spanish distance-learning courses, and one in which the course tutor has an indispensable role.

The information provided in this segment of our assessment system is the result of the annotations and observations made by the tutor on the student's performance with his or her group in the Round-up Exercises and End of Lesson Tasks. There exist other communication exercises, which we might refer to as 'free', which are not marked by the tutor and in which the student can practise the Spanish learned in real communicative exchanges.

To clarify further the features of the Round-up Exercise, we have drawn an example from those given to students in the course. In the exercise, the students are expected to use what they have learned in the relevant stage. In this particular case, they are asked to construct a character using the elements that come up on the screen (name, surname, age, nationality, occupation), and which they have been practising in the stage.

For the next step the student has to chat in real time with two companions and introduce the invented character to them. Similarly, the other students introduce him or her to their own invented characters.

The final step consists of jotting down the received information and sending it to the tutor to receive his or her comments.

A similar structure, but applied to the entire lesson, is the End of Lesson Task. The three Round-up Exercises function as gateway tasks to prepare the student to do the End of Lesson Task for each lesson.

Let us describe the exercises in the End of Lesson Task for Lesson 2. To begin with, the students have to make up a fictitious family with some of the

characters provided. Following this, they define a personality for each of them by giving them features. Finally, they have to contact another student and the working pair then exchange information on their respective families.

As before, sending the results of the exchange to the tutor to receive his or her advice and suggestions for improvement completes the task.

Course Tutor's Assessment

The student can access the Course Tutor's Assessment from the Assessment option in the Study Room.

This space should be visited frequently as, apart from the evaluation, the student is privy to a series of comments geared towards improving his or her performance.

If, once inside the Assessment System, the student chooses the Course Tutor's Assessment, he or she can find out the marks and improvement suggestions the tutor has entered for the Round-up Exercise of each stage, as well as the End of Lesson Task for each lesson.

The figure below shows the information available in the Course Tutor's Assessment.

Figure 6

The tutor provides an evaluation that reflects the main communication skills shown by students in these exercises. The rating scales on which the students' reports rest are outlined below.

Rating scales

Each of these templates evaluates a particular skill. This evaluation comprises a variable number of assessable areas (register, style, vocabulary, grammar, etc.). We have adapted the proposals presented in Alderson, Wall and

Claphan (1998) and the instructions for examiners of the Basic Diploma in Spanish as a Foreign Language (2001) for our own use.

The computer, in accordance with the areas evaluated by the tutor, automatically calculates the total mark in percentages. The student also receives suggestions for improvement.

In each assessable area there are three qualifications:

1. Excellent (>85%)
2. Satisfactory (>60% <85%)
3. Improvable (<60%)

WRITING SKILLS RATING SCALE

STYLE:
Respects the style of the text. Satisfactory register.
Respects to a limited degree the style of the text. Register not wholly consistent.
Does not respect the style of the text. Register inadequate.

STRUCTURE:
Content is clearly organised: connectors used correctly and punctuation satisfactory.
Content only slightly organised: connectors are neither sufficient in number nor used correctly.
Content disorganised. Connectors absent or used incorrectly.

GRAMMAR:
Almost all the grammatical structures for the level are used correctly.
Frequent errors in the grammatical structures for the level.
Abundance of basic grammatical errors.

VOCABULARY:
1. Satisfactory vocabulary, although occasionally imprecise.
2. Vocabulary frequently unsatisfactory. Vocabulary errors.
3. Vocabulary almost entirely imprecise, with many errors.

SPELLING:
Almost no spelling mistakes for the level.
Frequent spelling mistakes for the level.
Abundant spelling mistakes for the level.

Improvement suggestions:_____

COMMUNICATION SKILLS RATING SCALE

REGISTER:
1. Register almost always satisfactory.
2. Register often unsatisfactory.
3. Register unsatisfactory.

INTERACTION:
Maintains almost permanent communication with interlocutor. Answers satisfactorily.
Communication frequently hesitant. Frequently requests clarification.
Unable to maintain communication with interlocutor. Constant requests for clarification.

GRAMMAR:
Almost all the grammatical structures for the level are used correctly.
Frequent errors in the grammatical structures of the level.
Abundance of basic grammatical errors.

VOCABULARY:
1. Satisfactory vocabulary, although occasionally imprecise.
2. Vocabulary frequently unsatisfactory. Vocabulary errors.
3. Vocabulary almost entirely imprecise with many errors.

Improvement suggestions:_____

Finally, we should keep in mind that on our courses students always have access to their course tutors through the email, forum or noticeboard facilities. They are free to consult them at any time, in the knowledge that help will always be forthcoming.

Conclusions

The Assessment System was designed to complement the courses' distance-learning method and communicative approach. Basically, it provides students with progress input from two main sources:
1. From an automatic follow-up and self-correction system.
2. From a system of assessment by course tutors.

These two information systems were designed to help students make conscious, informed choices as to the appropriateness of certain strategies when learning a foreign language.

The first system provides information on students' achievements in different areas of language learning. It stores information from presentation, conceptualisation and practice exercises covering different linguistic and socio-cultural areas of Spanish. These exercises involve discrete-point items that can be computer-assessed. However, exercises requiring the production of a creative text, such as a composition, or communicative exchange with a partner, come up against the limitations of a computer system incapable of assessing them.

Therefore, the second system was designed to help overcome these limitations by the provision of tutors equipped to inform students of their performance in meaningful tasks. Here they have to activate and integrate knowledge to achieve different kinds of communicative goals, such as creating and receiving texts in chats, emails, etc., using all the skills necessary for communication.

The computer's capacity to store information is one of the main advantages of distance-learning, as it provides students with continuous follow-up information which they can consult at any time. By requiring students to store data and visit the automatic follow-up and self-correction system, we have tried to make them aware that their learning process is to a large extent their own responsibility.

The role of tutors is to offer specific improvement advice and help students focus not only on linguistic and socio-cultural elements, but on learning

strategies as well. This makes the assessment system consistent with the methodological principles of the *Cursos*, and its success depends on whether the material stored from assessable exercises (electronic mail, chat transcriptions, etc.) provides tutors with sufficient information to carry out a proper needs analysis for each student.

The creators of the Assessment System have also wished to establish a virtual study community by incorporating the work pace space. Here, individual students can compare their work pace with that of the group. In this way, a student who is falling behind can work with tutors and group companions to seek out the reasons and find ways of avoiding isolation. This is important if group exercises are to work.

In short, the success of the Assessment System depends largely on whether or not the weighing of tasks between the automatic system and the tutorial system, along with group work, gives students enough information to be able to identify their problems and take charge of their own studies.

In the future, we should focus on two pedagogical issues that still provide a challenge for creators of distance-learning systems. On the one hand, we must find out how to equip self-evaluation systems with the means of identifying and classifying any difficulties students might have in their learning strategies and style, through specific, automatic messages designed to help them think about and self-evaluate their work. On the other hand, to solve the problem of assessing oral performance, we need communication tools, such as chat systems and audio-conferencing, to be more reliable, more easily accessible and faster.

The Assessment System, which is constantly revised and updated by Instituto Cervantes staff, is implemented by a team of technicians at the Institute of International Economics at the University of Alicante. Revision is twofold: on the one hand, we check that the two elements of the Assessment System (student-based and tutor-based) work in an integrated and consistent manner. On the other hand, we use feedback from real groups of students studying the *Cursos*.

References

Arrarte, G., A. Duque, G. Hita, O. Juan, J. M. Luzón, I. Soria y J. I. Sánchez. Instituto Cervantes. 'Cursos de español a distancia a través de Internet. Una experiencia de innovación pedagógica del Instituto Cervantes'. Congreso Internacional de Informática Educativa 2000, Universidad Nacional de Educación a Distancia.

Bates, A. W. 1995. *Technology, Open Learning and Distance Education*. London: Routledge.

Chapelle, C. A. 2001. *Computer Applications in Second Language Acquisition. Foundations for Teaching Testing and Research.* Cambridge: Cambridge University Press.

Giovannini, A., E. Martín Peris, M. Rodríguez and T. Simón. 1996. *El proceso de aprendizaje*, Madrid: Edelsa.

Holec, Henri 1979. *Autonomie et apprentissage des langues étrangères*. Council of Europe, Modern Languages Project. Nancy: Hatier.

Fernández Pinto, J. 1999. 'Servicios telemáticos: impacto en función de sus características'. Cuadernos Cervantes, 23. Madrid.

Soria Pastor, I. y J. M. Luzón Encabo. Instituto Cervantes. 'Un sistema de seguimiento inteligente y evaluación tutorizada para la enseñanza y aprendizaje de segundas lenguas a distancia y a través de internet'. http://cvc.cervantes.es/obref/formacion_virtual/formacion_continua/-ines.htm, 2000.

Bibliography

Alderson, J. Charles, D. Wall and C. Clapham. 1998. *Exámenes de idiomas, elaboración y evaluación*. Madrid: Cambridge University Press.

Council of Europe. 'Modern Languages: Learning, Teaching, Assessment. A Common European Framework of Reference'. Strasbourg. http://culture.coe.fr/lanf/eng/eedu2.4.html 25/05/01.

Diplomas de Español como Lengua Extranjera. 2001. Examen para la obtención del Diploma Básico de Español como lengua extranjera del Ministerio de Educación y Cultura. D.B.E. 'Instrucciones para examinadores'. Ministerio de Educación, Cultura y Deporte, Instituto Cervantes, Universidad de Salamanca.

14 Certification of knowledge of the Catalan language and examiner training

Mònica Pereña and Lluís Ràfols
Direcció General de Política Lingüística,
Departament de Cultura, Generalitat de Catalunya

This paper consists of two parts. The first part describes the system of certificates of knowledge of the Catalan language resulting from the merger of the two systems in force until 2001.

The second part describes the distance-training system for examiners using a virtual training environment introduced by the Direcció General de Política Lingüística (DGPL). This system was used for the first time in 2001 for initial training and, from 2002 onwards, it has been used for the continuous training of examiners.

The certificate system for knowledge of the Catalan language

The Certificate System

The certificate system for knowledge of the Catalan language consists of five competence levels for accrediting general language knowledge:

- Certificate of Basic Level Catalan
- Certificate of Elementary Level Catalan
- Certificate of Intermediate Level Catalan
- Certificate of Proficiency Level Catalan
- Certificate of Higher Level Catalan

and four certificates for specific knowledge:

- Certificate of Knowledge of Administrative Language
- Certificate of Knowledge of Legal Language
- Certificate of Knowledge of Commercial Language
- Certificate of Skills for the Correction of Oral and Written Texts

The general language certificate levels correspond to ALTE's framework of five levels and those established by the Council of Europe in its *Common European Framework of Reference for Language Learning, Teaching and Assessment*, with the exception of the lowest level – Breakthrough Level – which has not been deemed necessary for Catalan.

The five new levels are a fusion of the two previous certificate systems:

- the certificates of the Junta Permanent de Català (Permanent Board for Catalan) – Certificate of Basic Oral Knowledge of the Catalan Language (Certificate A), Certificate of Elementary Oral and Written Knowledge of the Catalan Language (Certificate B), Certificate of Intermediate Oral and Written Knowledge of the Catalan Language (Certificate C) and Certificate of Higher Oral and Written Knowledge of the Catalan Language (Certificate D) – and:
- the International Certificate of Catalan (Basic Level, Threshold Level and Advanced Level).

Background

The certificates of the Permanent Board for Catalan were based on those of the Tribunal Permanent de Català (Permanent Committee for Catalan), a body created in 1934 by the Generalitat of Catalonia and chaired by Pompeu Fabra during the Republican Generalitat; the certificates were introduced following the recovery of the democratic institutions and self-government in 1979.

Since the Catalan language had been prohibited and excluded from the curriculum, Catalan citizens had been unable to learn this language at school and were thus unable to demonstrate their knowledge of it with academic certificates. Once the official nature of Catalonia's language had been recognised, the certificates of the Permanent Board for Catalan were introduced to create a system for evaluating knowledge of the Catalan language for these citizens.

The International Certificate of Catalan was created by a mandate of the Catalan Parliament in response to rising levels of Catalan language teaching abroad and the need for accrediting this knowledge; this increase was brought about by the growth of labour and academic mobility of EU citizens, within the context of integration, and by increased immigration from non-EU countries.

Over the last 20 years important social changes have been taking place in Catalonia, mainly due to three factors: the academic requirement of being able to use both official languages (Catalan and Spanish) normally and correctly by the end of compulsory education; legislative changes affecting the legal treatment of the Catalan language, and, finally, the organisation of government bodies. Moreover, the experience accumulated over almost two decades of assessing linguistic competence and the adaptation to new

language assessment methods have dictated a review of the current certificate system and the creation of a new regulation.

The new regulation

This new regulation places the Department for Culture's Directorate General for Language Policy in charge of holding and administering Catalan language examinations and for issuing the new certificates. This has led to the disappearance of the Permanent Board for Catalan, which had previously been responsible for awarding certificates. Its members were appointed by the Minister for Culture from individuals qualified for teaching Catalan who were often members of bodies that filled the gap produced when the activity of the Permanent Committee for Catalan was suspended. The International Certificate of Catalan examinations will also be withdrawn.

In accordance with the decree, the task of co-ordinating the bodies responsible for assessing Catalan knowledge in other Catalan-speaking countries is delegated to the Department for Culture.

The certificates regulated by this new decree are official Catalan language certificates and certificates of reference. The decree also provides for the possibility of creating qualifications, diplomas and certificates equivalent to those regulated by this decree, by order of the Minister for Culture.

These equivalents are created with the single aim of permitting the following to accredit their knowledge of the Catalan language: candidates who want to enter the Catalan civil service or civil servants who want to take part in promotions or mobility competitions in their respective local governments. These candidates will thus be exempt from the compulsory Catalan language examinations of recruitment schemes.

The main changes brought about by this new regulation are the disappearance of the Permanent Board for Catalan, a new description of levels of language knowledge, and a change in the examination structure and sections for each certificate presented as an attachment to the decree.

Broadly speaking, the first three levels – Basic, Elementary and Intermediate – are the three levels of the old International Certificate of Catalan.

The Proficiency Level puts more emphasis on the assessment of the following aspects than did the old Certificate of Intermediate Oral and Written Knowledge (Level C): ability to produce both oral and written texts; command of the linguistic system by applying general and specific rules, and rules of exception; choice and usage of parts of speech in spoken and written contexts, taking into consideration syntax and semantics; ability to produce grammatically correct oral structures in Catalan; formation of words by applying the most common mechanisms of derivation and composition, and knowledge of the exact meaning of words and phrases.

It also aims to assess the candidate's ability to express him- or herself

adequately with a sufficient level of correctness in communicative situations based on social and occupational contexts (one-way communication) with an average level of formality by making a two-minute presentation on a general theme and by reading a text aloud.

The main differences as regards the previous examination are in the reading comprehension section, where the length of texts has been extended, in the number of items and the distribution of assessment objectives in different types of text, and in the oral expression section, where marking is carried out by two examiners. A further exercise has been added and more concepts and grading have been introduced for assessment purposes.

We believe that these changes have increased coherence between the assessment objectives and the examination design because there is more context and stimulation, the grammar section items are mixed and spelling is assessed indirectly, while greater importance is given to syntax and vocabulary.

The main aim of the higher level certificate is, firstly, to evaluate whether the candidate has sufficient general linguistic command to communicate adequately and correctly in any communicative situation, particularly those that require more formal language. Secondly, it aims to assess whether the candidate is capable of analysing matters and concepts related to correct use of the Catalan linguistic system.

This change in the Catalan knowledge certificate system has thus led to:

1. The creation of a single valid certificate system for individuals requiring accreditation of their knowledge of the Catalan language, regardless of whether they are Catalan-speakers, long-term residents of a Catalan-speaking area, or individuals who learn Catalan through initial contact with this language (either abroad or during a stay in Catalonia).
2. The adjustment and increased precision of the grading of levels of knowledge of the Catalan language and adaptation of these to the Council of Europe's framework of reference for the assessment of modern languages as well as that of ALTE (Association of Language Testers in Europe).
3. The adaptation of Catalan language-certificate examinations to new linguistic assessment criteria, with modifications in the types of exercise and grading scales of the process.
4. A decree-level standard regulating the certificates that officially accredit the adult population's knowledge of Catalan and the examinations for obtaining these, and the detailing of the structure and content of said examinations.

In order to adapt the task of examiners of certificate examinations to the changes in this new system, we designed the virtual training environment described below.

Virtual training of examiners

Background

The Directorate General for Language Policy (DGPL) has introduced a scheme for training examiners using a virtual environment. This system was used for the first time in 2001 for initial training of examiners of a range of certificates. In 2002, it was also used for continuous training.

The decision to opt for distance training using a virtual environment was influenced by:

a) the updating of the Certificate of Proficiency, which created the need to retrain the entire group of examiners of this certificate in applying the new criteria and assessment grading scales. This group numbered some 337 individuals in total
b) the need for a distance method of training examiners, in order to increase the possibilities of administering certificates of Catalan in areas outside Catalonia without examiners in training or their tutors having to travel. A total of 65 individuals had to be trained for Elementary Level and 19 for Basic Level.

The training of Proficiency Certificate examiners was performed in March and April of 2001. Nine months before this, examiners were informed that training would involve virtual learning. A questionnaire was carried out beforehand to find out the group's level of computer equipment: in the summer of 1999, approximately 60% of examiners already had a multimedia PC with an Internet connection, and 20% said that they would have this equipment within nine months' time.

Training for the Basic and Elementary Levels of the International Certificate of Catalan was performed during February, March and April of 2001. Nine months beforehand, training candidates were informed that this system would be used. The majority were *lectors* in universities all over Europe and the availability of hardware varied according to country and university. As a result, solutions had to be found in the virtual environment to take into account the circumstances of certain areas.

The environment

The advantages of a virtual training environment for preparing examiners are as follows:

a) It allows the training of large, geographically-dispersed groups of individuals without their having to travel.
b) The number of hours of training can be increased because it does away with the time limits of on-site training (in our case, 1 day: 7–8 hours training) with the possibility of performing between 15 and 20 hours' training.
c) It enables examiners in training to work at their own pace (within the set

schedule) and go over material as often as necessary. This is particularly important for oral expression: examiners can watch the video of the candidate and analyse it in detail with the assessment grading scales. Clearly, this is not a real examination situation, but as a training situation it allows examiners to study carefully how assessment criteria are used.

d) It allows system administrators to see what is happening in each classroom and to unify the working criteria of tutors quickly and almost instantaneously; information on the progress of training can be collated quickly and opportune decisions can be taken during the training period itself.

Obviously, this method of training is conditioned by certain factors. Firstly, the correct hardware and software are required: examiners in training must have a multimedia PC, modem and Internet connection and certain standard programs: Word, Acrobat Reader and a multimedia player. Secondly, they need to be fairly familiar with new information and communication technologies: basically, they need to know how to use hypertext browsers and email.

Another conditioning factor is the fact that tutors of the virtual environment must also be trained, because, from a pedagogical point of view, there are numerous aspects that do not exist in classroom training which need to be taken into consideration. With this in mind, a ten-hour training course in virtual learning environments was set up for trainers.

The virtual training environment

We opted to hire the virtual campus of the Universitat Oberta de Catalunya (http://www.uoc.edu) because of this University's experience in virtual distance learning. The visual aspect of the campus was adapted to the corporate image of the Department for Culture of the Generalitat of Catalonia. There was also a telephone service for users (both tutors and examiners in training) to ask questions about computing equipment.

This environment basically uses an asynchronous model of communication. Communication is deferred and the individuals involved do not need to be connected to the environment at the same time (which would be otherwise unfeasible). However, a chat function allows on-line conversation.

The virtual environment contains the following areas:

TRAINING area: classrooms

The most important elements of each classroom are the mailboxes and the file area.

Mailboxes: there are three mailboxes, each with different access privileges.

Noticeboard: all members of the classroom can read messages, but only the tutor can write them.
Forum: everyone can read and write messages; this area is designed for informal communication.
Debate: this has the same features as the forum, but is reserved for more formal discussion about matters specifically related to training.
File area: here, the tutor can add all the documents needed for work or reference.

ROOMS area: staff room
This area is structured like a classroom, but access is restricted to the tutors of each classroom and the noticeboard is administered by the DGPL staff in charge of training. The noticeboard can provide instructions, guidance and performance criteria for tutors.

Access to other Internet resources: Catalan language website and Internet search engines
These two options offer access to on-line Catalan language resources of the Department for Culture and to different search engines.

DGPL Area: organisation notice board
This area is the notice board of the organisation. It can be used, for example, to offer examiners in training an institutional welcome or for any other type of institutional notice.

Personal mailbox
Each member of the virtual community has a mailbox.

Other functions: Directory, Who is connected, Preferences
These options enable us to look for a specific person in the virtual community (Directory), find out who is connected to the environment at a particular time (Who is connected) and to change the settings of various options of the personal mailbox.

Training

Training material

Material was distributed on CD-ROM because a video was included and this would have made downloading from the Internet difficult.

The initial training material for Proficiency certificate examiners – which involves oral and written assessment – is described briefly below. For Basic and Elementary Levels, training is only available for assessment of oral expression, because written expression is marked centrally and correctors are trained in on-site sessions.

The general browser bar for all the material is located in the top part of the screen. From here, the user can access the following:

Learning guide (contains the work plan and training guidance)
Global view of the examination
Area 1. Written expression
This block contains a presentation of the area and a different section for each concept that requires assessment (order and register, cohesion, variation and correction); for each concept, there is a sample assessed text and a self-correcting partial practical. Lastly, there is a section with sample texts and overall practicals in this area.
Area 2. Oral comprehension
This block contains a presentation of the area.
Area 3. Grammar and vocabulary
This block contains a presentation of the area.
Area 4. Oral expression
This block contains a presentation of the area and the following sections: examination administration, assessment of exercises 1, 2 and 3; there is an assessed sample for each exercise and a self-correcting partial practical. Lastly, there are global practicals for this area.
Briefcase. Examiner documents: assessment grades and other documents.

There are two global practicals, for both written and oral expression. Examiners in training have to perform the first global practical and send it to their tutor, who then marks it and returns it to the examiner with comments on the level of concordance with the assessment set for this particular practical. These comments allow the examiner to adjust to the set criteria in preparation for the second global practical.

For written expression practicals, a Word template was used to simulate the marking of a handwritten examination on-screen using a toolbar that reproduced the signs examiners have to use to mark the real examination.

This graphic marking tool is one of the elements that need to be revised as it caused the most problems and was evaluated negatively because it was not practical enough. It was used particularly in initial training because of the need to train examiners in the use of standardised marking signs. Use of this tool has been avoided in continuous training as it is assumed that the criteria for use of signs is already known.

Work Schedule

The length of training is particularly important when planning this type of method. The virtual environment imposes its own work pace which requires the training period to be extended. At least one month is required to carry out 12–15 hours' training. Communication is asynchronous, which means that a message sent at 11 p.m. one day may not be read by the rest of the group until 10 a.m. the next. If the message is an instruction from the tutor or the answer to an important question about the use of assessment criteria, tutors cannot expect examiners to follow this instruction or heed this piece of advice until the following day. This clearly affects the creation of a training work schedule.

Examiners in training

Before starting training, examiners were given two days to familiarise themselves with the virtual environment, material and communication system.

For training of the group in 2001, the full potential of the virtual environment as a communication tool was not exploited because the group was new to this type of system. For example, the Debate area allows group discussion on the assessment of a sample of written and oral examinations, which is a common practice in on-site sessions. It could therefore be said that the training experience carried out in 2001 was a form of individual tutored learning that did not exploit the possibilities of interaction and group work in virtual training environments.

For continuous training in 2002, we assumed that the examiners were already familiar with the virtual environment and the debate element became the main focus of training in order to adjust the criteria of examiners to those of the DGPL. The training used guided debates on aspects that the DGPL considered should be brought to the attention of examiners for reflection and analysis.

Although both tutors and examiners regarded this option favourably, tutors had to make a great deal of effort to encourage examiners to take part in the virtual debates proposed with varying results. Examiners were informed that the result of the training depended on their participation in debates and that they had to take part in each debate at least once to inform the rest of the group about their opinion on the proposed topics. The overall success of the debate and the number and quality of contributions varied from group to group; it is hard to discern the reasons for this variation (tutors and examiners were allocated to groups at random), but it could be that certain tutors or examiners had the ability to encourage participation (a sense of humour, use of the forum for informal discussion with examiners about cinema, books and other topics not related to training are some elements that help to break the ice and encourage participation).

Examiners in continuous training were assessed on the basis of the amount and quality of their contributions. For example, an examiner who participates as little as possible (a single contribution per debate) with poorly reasoned contributions will be awarded a low mark. Depending on the case in question, the DGPL may deem the candidate to have failed the training and to be thus unfit to examine.

Tutors

As we said earlier, tutors underwent initial training on learning in virtual environments. However, their relationship with the project began in the phase prior to the development of training contents: in order to involve tutors in the process and to ensure a solid knowledge of the agreed assessment criteria and the arguments required to provide support, they participated actively in the analysis and assessment of examination samples to be presented to examiners in training.

The task of these tutors was mainly to guide examiners in training when applying assessment grading scales and to return their practicals with the appropriate comments and advice. To allow tutors to do so correctly, a low tutor/examiner ratio was used: 12–13 examiners in training for every tutor. Moreover, the technical staff of the DGPL responsible for each exam and for managing the environment provided support and advice to tutors throughout the training period regarding any questions or problems that arose. The ROOMS area of the environment, which acts as a staff room, enables this support to be given almost instantaneously and ensures unified criteria for resolving the questions of examiners about training content that were not covered when the training material was validated, and for resolving problems about the management of the virtual environment.

When training was completed, tutors had to prepare an assessment report for each examiner in training on how suitable they would be for the task of examining. On the basis of this report, the DGPL made its final decision as to whether or not that individual would be fit to examine.

Table 1 illustrates overall data on the number of individuals involved in this training system.

Table 1 Overall data on individuals involved in the training system

Year	Examiners in initial training	Examiners in continuous training	Tutors
2001	421	—	42
2002	42	236	31

Assessment of the experience

The opinions of examiners and tutors were obtained by using a questionnaire that dealt with a range of aspects. The answers to two questions from this questionnaire provided us with relevant information about how examiners and tutors viewed this type of training system.

In the 2001 edition, 66% of examiners considered this training system to be quite or very adequate (on a scale of not at all/hardly/quite/very) and 78% considered that they had taken little or no advantage of the environment's possibilities for communication (asking questions …) with the tutor and other examiners in training: this assessment is logical, seeing that the DGPL had decided not to take advantage of the possibilities (for interaction and group work) of virtual training environments and to limit the new system to individual tutored learning.

However, in the 2002 edition, which included debates guided by the tutor

to encourage discussion and reflection among examiners, 73% of examiners considered that they had benefited a fair amount or a great deal from the possibilities for communication in the environment; 95% also considered that the system of work based on debate and exchange of opinions was quite adequate or very adequate for revising criteria and addressing queries. Significantly, 65% of examiners considered that they rated the usefulness of this training system more highly than that of the 2001 edition.

Assessment of tutors

The assessment of tutors is generally similar to that of examiners. They commented on the need for improvement in certain areas of their task. For example, in initial training of the Certificate of Proficiency Level, they considered that the support documentation provided by the DGPL to help them with their task was too much and needed to be simplified; they also expressed difficulties in meeting the set schedule.

The continuous training of 2002 was also rated positively overall. However, in general, tutors pointed out the need for a strategy to ensure that the debates proposed for the examiners benefited from as much participation as possible because tutors have to assess the abilities of each examiner on the basis of their contributions: it is difficult to assess an examiner who has not contributed or who has made poorly-reasoned contributions, other than by failing him or her for this reason.

The overall impression has been positive in the sense that the virtual training environment is a useful tool for training a group such as examiners, which is normally a geographically dispersed group. Clearly, there are certain aspects that need to be improved, such as the written-expression graphical marking tool, and some aspects of training material need to be made user-friendly; these changes, however, relate to the design of the material and its adaptation to the multimedia context and not to the method itself.

Other aspects also need to be studied in detail if we are fully to exploit the possibilities of virtual training methods. The introduction of debates in the continuous training of 2002 is a step in the right direction. However, we need to search for new options and training strategies for group work in virtual environments to create even more productive and useful situations for discussion and reflection.

15 CNaVT: A more functional approach. Principles and construction of a profile-related examination system

Piet van Avermaet (KU Leuven), José Bakx (KUN), Frans van der Slik (KUN) and Philippe Vangeneugden (KU Leuven)
CNaVT (Dutch as a Foreign Language Certificate)
www.cnavt.org

Introduction

CNaVT (Certificaat Nederlands als Vreemde Taal – Dutch as a Foreign Language Certificate) is a government-subsidised, non-profit organisation that was founded in 1975 and was placed under the auspices of the Nederlandse Taalunie (NTU) in 1985. Since 1999, CNaVT has been affiliated with the Centre for Language and Migration (CTM) of the Catholic University of Leuven (KU Leuven) in Belgium and the University Language Centre (UTN) at the University of Nijmegen (KUN) in the Netherlands. The NTU asked Nijmegen and Leuven to develop a more functional certification structure, with new proficiency tests in Dutch as a foreign language.

In this article we will focus first of all on the paradigm shift that led to this new certification structure, and on the process of central test construction that we went through. This construction process started with a needs analysis, followed by a discussion on profile selection. We then present the selected profiles that formed the basis of the new certification structure that is presented in the following section. In this section, we also clarify the relationship of the new certification structure with the development of a test bank, which was another part of the NTU assignment. In the final section we present the way the exact content and difficulty of the profiles is described.

For more information on the CNaVT examinations, please refer to the websites of ALTE (www.alte.org/members/dutch) and CNaVT (www.cnavt.-org).

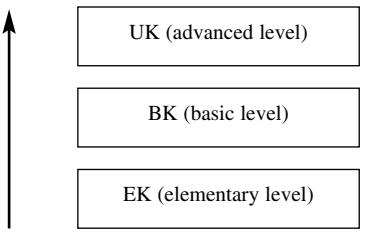
A paradigm shift

Learners aspire to a command of the Dutch language for a variety of reasons. They might be going to work for a company where Dutch is the official language; they may have family or friends in Belgium or the Netherlands, or they may want to understand legal documents in Dutch. In short, most learners want to learn and acquire a command of the Dutch language in order to be able to function in certain areas of society. This implies that not every individual has the same language needs, and thus diverse needs can be observed.

In many cases the individual and/or society needs proof of the desired or required language proficiency (Humblet and van Avermaet 1995). Certificates are relevant when a clearly defined form of language proficiency is a prerequisite for admittance to a part of society or when it needs to be established that somebody can function in a certain cluster of situations. This diversity, variation and need for contextualisation has to be taken into account when developing a new certification structure.

The old Dutch as a Foreign Language examination system did not start from this perspective. It can be presented as a vertical progressive model testing general language proficiency (see Figure 1). A distinction was made between the examination levels, but not between the contexts in which candidates wanted to function. The exams did not take into account the language variation that can be observed in different communicative situations. It was a vertical progressive system in the sense that many candidates first took the lowest level (EK) and then often climbed the ladder to the highest, paying for every rung on the ladder.

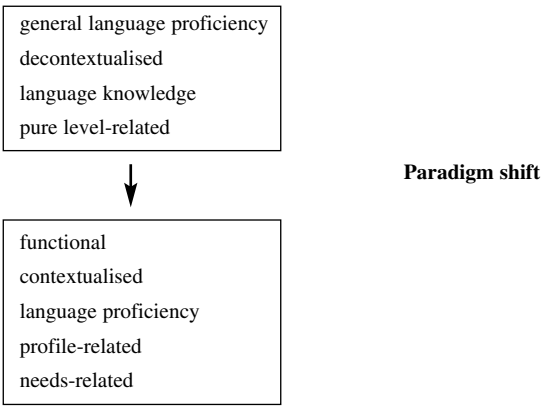
Figure 1 Old certification structure



The new CNaVT is intended to reflect the needs of the target group more closely. The new certificate should demonstrate that a sufficient command of the language has been acquired in order for the learner to function in the situations and domains in which he or she wishes to use the Dutch language. To have a command of the language means that one is able to use it for all kinds of communicative purposes. In other words: the exams will have to test

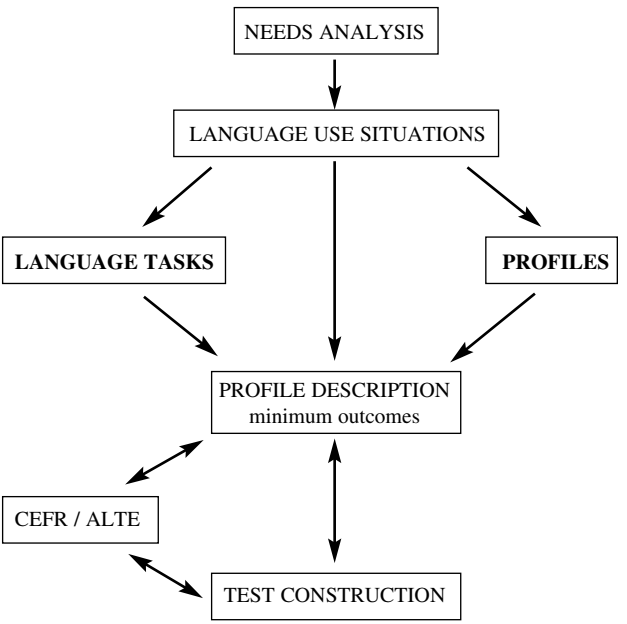
functional language proficiency. Figure 2 summarises the paradigm shift that was necessary to meet these points of departure.

Figure 2 Paradigm shift: old to new



In the process we went through in order to complete this new certification system, an underlying structure can be recognised. This structure is shown in Figure 3 and will be discussed in detail in the rest of this article.

Figure 3 Structure of test construction process



Needs analysis

All exams in the new examination system should have accreditation: a recognition of the social relevance of the exam and a guarantee to both the individual and society. This implies that the relation between what is tested and the areas in which one wants to function is very important. The decisive factor in gaining this accreditation is the degree to which the constructs assessed in the examinations match the existing language proficiency needs.

The first step that had to be taken was to gain insight into the candidates' needs. We tried to assess the needs of students of Dutch all over the world through a written questionnaire that was sent to a representative sample of students and teachers of Dutch as a foreign language. A part of the questionnaire consisted of spheres in which the student, using the Dutch language, might want to function. Examples were 'law study', 'work in business' or 'living in the Netherlands or Flanders'. In addition, space was offered for specific extra spheres. A second part of the questionnaire contained a list of 30 concrete situations in which students might want to use the Dutch language after they had completed their course. This list was not exhaustive, but consisted of a range of diverse situations within different domains. Teachers were asked to check – based on their experiences – whether these domains and situations were important, less important or unimportant for their students. The students were asked to do the same.

The analysis showed that the questions relating to situations had the most interesting results. We will therefore only present the results concerning the situations. They were classified by means of principal factor analysis, in order to detect underlying concepts. We used standard procedures such as the Scree test to perform these analyses.

The analyses of the teachers' data resulted in four separate dimensions that were preliminarily designated as 'Business contacts' (clustering situations such as, for example, having a business meeting, making a business telephone call, writing an article), 'Social contacts' (e.g. buying in a shop, calling relatives, contacting school teachers, following Dutch broadcasting), 'Study' (e.g. taking an exam, following a course) and 'Tourism' (e.g. making a hotel reservation, being shown around, reading tourist brochures). Social contacts and Tourism were perceived as the most important areas. The outcomes for the students were partly in line with the teachers' data. Students made no distinction between Social contacts and Tourism, but they too perceived these to be the main areas in which they intended to use Dutch, again followed by Study and Business contacts.

Qualitative phase and profile selection

The quantitative analysis was followed by a qualitative discussion in which we consulted the advisory board as well as other key informants in the field of Dutch as a Foreign Language. In this phase the following considerations were made, in order to establish profiles that are as relevant as possible for substantial groups of learners of Dutch as a Foreign Language.

First, we had to interpret the factors revealed by the statistical analysis. The labels given in the previous phase were only preliminary. Here we dealt with questions such as how to describe a certain cluster of situations in a meaningful way. In order to achieve an accreditation for the examinations it is necessary for a cluster to be clearly communicated to both candidates and society.

Another question was whether these clusters are relevant as a basis for certification. The importance of the selected situations for students was clear from the survey. But possibly not for every cluster is it relevant for society to certify the language proficiency needed. For example, our PTIT profile (see Table 1) – describing Dutch language proficiency in the tourist and informal domain that was found to be relevant to our target group – could be considered less important for society to certify. However, this profile was retained because of the motivating effect of its certification, communicated to us by certain key informants.

Finally, it proved necessary to limit the number of profiles to four in order not to end up with an exam for every candidate. Taking into account the diversity in needs and context, more specific exams would be interesting from a theoretical point of view, because the degree of contextualisation and variation would be very large. However, for economic and administrative reasons, it was decided to opt for four more generic profiles, which nevertheless take contextualisation and diversity into account.

Four profiles

The above quantitative and qualitative analyses led to the selection of the four profiles presented in Table 1. The acronyms do not match the English labels but rather the official Dutch labels.

After the selection of these profiles was completed, it was clear that they were very much in line with the domains identified in the Common European Framework of reference for languages (Council of Europe 2001) and by the Association of Language Testers in Europe (2001), namely social/tourist, work and study.

Table 1 The four resulting profiles

Profile academic language proficiency (PAT) This profile covers the language proficiency needed to function in Dutch at an academic level. The profile is suitable for those wanting to demonstrate that they can deal with texts both orally and in writing at the academic level. Examples are reading (scientific) literature, giving a presentation, consulting resources, etc.
Profile professional language proficiency (PPT) This profile covers the language proficiency needed to function in Dutch at the professional level. The profile is suitable for those wanting to demonstrate that they can function at the level of middle management. Examples are writing business letters, running a business, attending meetings, etc.
Profile social language proficiency (PMT) This profile covers the language proficiency needed to function in society in Dutch. The profile is suitable for those wanting to settle in a Dutch-speaking area for a short or a long period or for those who are interested in Dutch language and culture,watching TV programmes in Dutch, reading Dutch newspapers, opening a bank account, conversing about current affairs, etc.
Profile tourist and informal language proficiency (PTIT) This profile covers the language proficiency needed to function at a basic level in Dutch. The profile is suitable for those wanting to maintain social contacts with Dutch family or friends or for those wanting to demonstrate that they can manage as a tourist in a Dutch-speaking area. Examples are making a hotel reservation, carrying on an informal telephone conversation with a family member, reading a tourist brochure, etc.

New certification structure and test bank

These four profiles form the basis for the new certificate structure. They are to replace the old certificate levels of Elementary, Basic and Advanced. Of course there will be a difference in language proficiency level requirements for the profile examinations. But the profiles do not constitute a linear hierarchical model, in the sense that a candidate first takes a PTIT- and PMT-exam and finally a PPT-exam, when he or she actually wants to function in a professional context. The name and graphic representation of the profiles (see Figure 4) try to make clear that candidates do not have to take the exam for PPT first to be able to take the PAT-exam, etc. Candidates should choose the profile that contains the situations and spheres in which they want to function.

A profile is seen as a level of competence or a final attainment level: the candidate has a sufficient command of the Dutch language to be able to function in the domain of his or her choice. In this line of reasoning, in-between levels are not relevant for certification because they do not demonstrate whether someone can function in a certain domain. (E.g. it is not relevant to an employer to know that a candidate is almost – but not quite – proficient enough to function in his or her company.)

On the other hand, during their language course, learners (and their teachers) would like to know where they stand and what they have learned. Therefore the construction of a database of tests – which was the second part of the assignment by the Dutch Language Union (NTU) – is important.

The test bank is intended to be a service for the teachers of Dutch as a Foreign Language. Its aim is to make an inventory of existing tests, to make them available to teachers (by means of a web-based search system) and to stimulate teachers to exchange their tests. Three different types of test will be put in the bank. In the first place teacher-made tests will be included. These tests have been developed by the teachers themselves and are often used in practice. In addition to these teacher-made tests, there will be space for recognised tests – such as the old CNaVT exams, for instance. Thirdly, the project team is taking the initiative to develop tests or stimulate and supervise the development of tests that fail to show up in the test database.

The test bank has a specific and very important goal that complements the centrally administered official CNaVT examinations. The tests in the bank are not aimed at the certification of a final attainment level and have no accreditation. Their aim is rather to guide the learning process. They offer teachers the possibility of assessing the level of their students' language proficiency on their way to the Dutch as a Foreign Language exams. Therefore there is specific particular provision for tests situated at the in-between levels. The tests in the bank will be described according to a number of general parameters, such as level, profile and tested skills (reading, speaking, grammar…). They will be related to the CEFR and ALTE levels where possible.

Profile description

The determination of final attainment levels or outcomes is a necessary next phase in the development of profile tests. The outcomes describe what people should be able to do with Dutch – and at which level – in order to function within a certain profile.

A first step in this phase was to make an inventory of the language-use situations that are relevant for each profile. We took as a starting point the situations that were in the needs analysis questionnaire. A second step was to look at the different possible language tasks people have to be able to fulfil in these language-use situations. For this, inspiration was found in Coumou *et al.* (1987).

In the enormous list of language tasks that resulted, a large amount of overlap could be observed. The language tasks that showed overlap were then clustered.

The next step was the description of the exact difficulty of each of the selected language tasks. For this we used a set of parameters that were inspired

Table 2 Example of some language-use situations and language tasks for profile PTIT

Language use situations	Language tasks
Making a hotel reservation by telephone	• understand/ask questions • answer questions • express requests and wishes • understand simple instructions • understand messages • understand/use occasional expressions
Reading a tourist brochure	• understand and select relevant data from informative texts
Asking/understanding a route description	• understand/ask questions • answer questions • understand simple instructions • understand/express messages • understand/use occasional expressions etc.

by Cucchiarini and Jaspaert (1996), the Common European Framework of Reference (Council of Europe 2001), the framework of reference for Dutch as a Second Language (Coördinatie-eenheid Prove 1996) and the Modern Languages Training Profiles developed by the Department of Educational Development of the Ministry of the Flemish Community (Dienst voor Onderwijsontwikkeling 2001).

As an illustration we present an excerpt from the description of the PTIT profile (see Table 3).

Table 3 Excerpt from detailed description of PTIT profile (part: listening)

Profile for Tourist and Informal Language Proficiency (PTIT)

1. LISTENING

LANGUAGE TASK	mastery level	text type	sender	register
Input				
understanding requests, wishes, complaints, ...	descriptive	informative, persuasive	unknown/ known	informal/ formal
understanding instructions	descriptive	prescriptive	unknown/ known	informal/ formal
understanding messages	descriptive	informative, persuasive	unknown/ known	informal/ formal
understanding occasional expressions	descriptive	informative	unknown/ known	informal/ formal

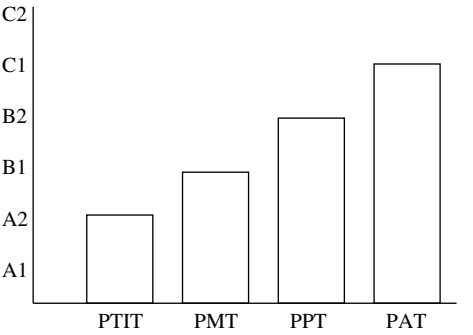
TEXT CHARACTERISTICS	*Input*
Vocabulary	Standard language is used. Words, formulations and idiomatic expressions are very common. Standard formulations occur frequently.
Formation of words and sentences	Mostly short sentences occur.
Pronunciation	The text is spoken with a standard accent. The speaker has a clear intonation and articulates well.
Speaking rate	The speaking rate is low.
Coherence	The text structure is clear. Texts are short and simply structured. Relations between sentences are indicated by frequent connecting words.
Content	The texts usually concern concrete topics.

MINIMUM FINAL OUTCOMES

- The candidate can determine the main message in requests, wishes or complaints (e.g. a request made by a hotel owner to make less noise during the evening hours).
- The candidate can determine the most important information in instructions (e.g. instructions given by a police officer during parking).
- The candidate can select relevant information from messages heard in everyday situations (e.g. a route description, a personal description, a guided tour).
- The candidate can recognise the most common conversational routines that arise at certain occasions and react to them appropriately (e.g. birthday congratulations).

Only after the detailed description of the four profiles is completed will a solid comparison be possible with the Common European Framework of Reference (CEFR 2001) and with the levels that were described by the Association of Language Testers in Europe (ALTE 2001). Figure 4 preliminarily situates the CNaVT profiles within the CEFR level framework, based on the profile descriptions as far as they are developed at this moment.

Figure 4 Profile-related certification structure



The CEFR functioned as a point of reference during the process, which seems to us the best way of using these European frameworks. We did not start from the CEFR level descriptions because the major focus points in renewing the CNaVT were the candidates' needs, diversity, functionality and contextualisation.

References

Association of Language Testers in Europe. 2001. *The ALTE framework. A Common European Level System*. Cambridge: Cambridge University Press.

Coördinatie-eenheid Prove (red.) 1996. *Eindtermen Educatie*. Amersfoort: Prove.

Council of Europe, Modern Languages Division. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Coumou W., *et al*. 1987. *Over de Drempel naar Sociale Redzaamheid*. Utrecht: Nederlands Centrum Buitenlanders.

Cucchiarini, C. and K. Jaspaert. 1996. Tien voor taal? Toetsen van taalvaardigheid. In VON-Werkgroep NT2 (eds.) *Taalcahiers: Taakgericht onderwijs: een taalonmogelijke taak*? Antwerpen: Plantyn.

Dienst voor Onderwijsontwikkeling. 2001. *Adult Education. Modern Languages Training Profiles*. Brussels: Ministry of the Flemish Community, Department of Educational Development.

Humblet, I., and P. van Avermaet. 1995. De tolerantie van Vlamingen ten aanzien van het Nederlands van niet-Nederlandstaligen. In E. Huls. and J. Klatter-Folmer (eds.) 1995. *Artikelen van de Tweede Sociolinguïstische Conferentie*. Delft: Eburon.

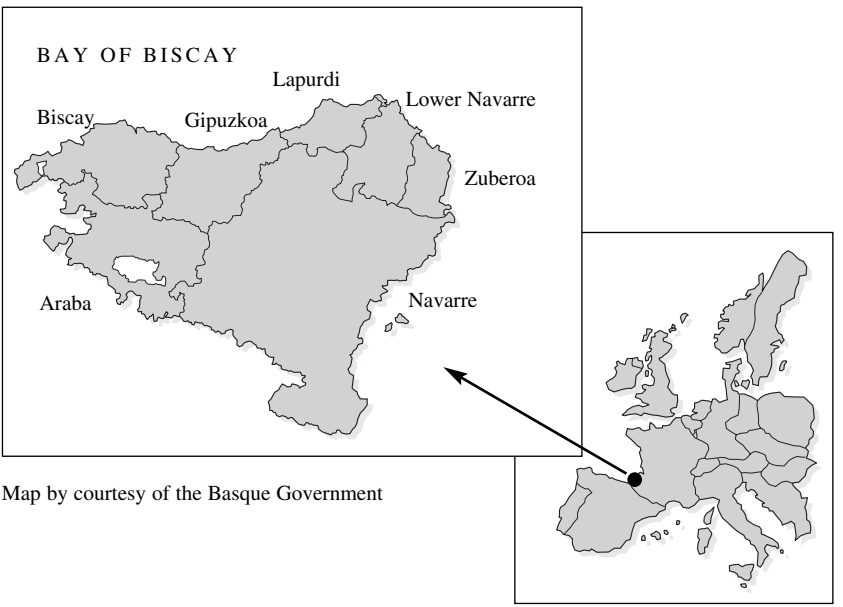
16 Language tests in Basque

Nicholas Gardner
Department of Culture/Kultura Saila
Government of the Basque Country/Eusko Jaurlaritza
Spain

Basque in the Basque country

Before discussing Basque language examinations in themselves, some scene-setting seems to be necessary, as the present status of Basque is little known. French, Spanish and Basque terms for 'the Basque Country' are ambiguous because they do not necessarily refer to the same geographical area, even within a single language. We are going to consider an older definition, an area of land including part of south-western France and an area several times larger in Spain in the bottom right-hand corner of the Bay of Biscay.

Figure 1 The Basque country in Europe



Map by courtesy of the Basque Government

Speakers of Basque in the area outlined constitute around a quarter of the present population of nearly three million. Practically all are bilingual in Spanish or French. The area was romanised and hence latinised to varying degrees: indeed, it seems Basque only survived because the Roman Empire gradually collapsed, preventing full romanisation. In the Middle Ages, Basque lost ground in some areas and gained in others. As far as records tell, it has always coexisted alongside other languages. A stable diglossic arrangement lasted until at least the 18th century. Thereafter, the story until very recently has largely been one of language loss: in southernmost rural areas initially, then (in the second half of the 19th century) in newly industrialised Spanish coastal areas, with intergenerational language transmission finally breaking down in the French Basque Country in the 1960s.

A first major attempt at reversing the language shift was carried out primarily in Spain from the beginning of the 20th century: it came to an abrupt end when Franco's rebel troops occupied Bilbao in 1937 during the Spanish Civil War. The second attempt, the ultimate source of the present language examinations, began in the mid-fifties in Spain. The initial phase – dependent on the sole forces of language loyalists – lasted until the end of the 1970s, when the present institutional phase began.

Figure 2

Map by courtesy of the Basque Government

Language policy-makers in the Basque Country vary considerably in the three main administrative areas indicated on the map: the French Basque Country, subject to the laws of France and the decisions of the *Conseil*

Général des Pyrénées Atlantiques; Spanish Navarre, subject to Spanish law and the decisions of its own regional government; and, finally, the Basque Autonomous Community (BAC), equally subject to Spanish law and the decisions of *its* regional government. The relative weight of the decisions of the three regional governments is very different, however, as over 80% of all Basque speakers live in the BAC. Discussion of examinations will centre precisely on the BAC.

Learners of Basque

Who learns Basque? Both native and non-native speakers. Most young native speakers are now schooled in Basque. However, as in many less-used languages, the term 'native speaker' is not always very helpful: such people have a very variable degree of command of the language, ranging from a literate, educated standard through competent oral command with limited literacy skills to minimal oral command, competent say in the home and with friends, but with major difficulties in any more formal register, and with limited reading ability. In addition, there are many second-language learners particularly from Spanish-speaking families. In any given year over 300,000 schoolchildren now receive Basque language lessons and a fair proportion also receive part or all of their education through the medium of Basque. To this total must be added a rather more modest number of university students studying at least part of their degree course through the medium of Basque and around 45,000 adults receiving Basque lessons.

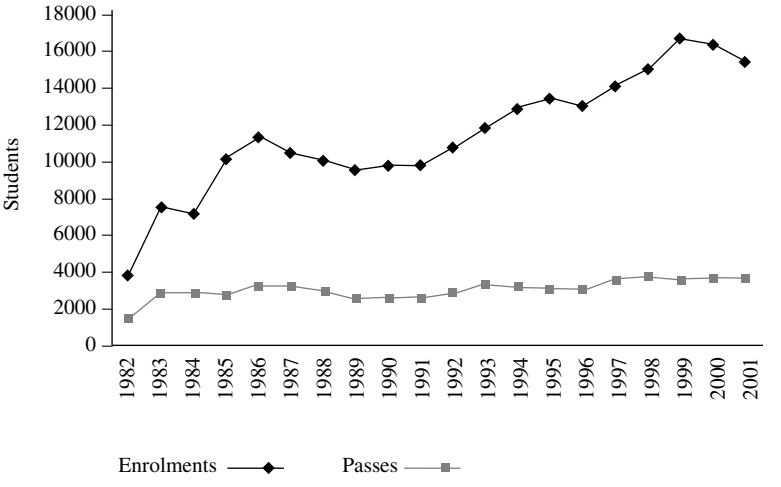
Motives are varied, but can conveniently be summarised under two headings. For some the objective is developing and obtaining proof of one's Basque language ability as a matter of pride in a cherished language; for others the need for Basque is more instrumental, closely linked to the belief – sometimes more imagined than real – that knowledge of Basque will improve their options in the job market.

A number of examinations are available to candidates: all have to cater for all comers, but the EGA examination run by the Basque Government is by far the most popular at its level (C1/ALTE level 5). The following graph shows enrolments and passes since its creation.

The rapid growth in enrolments in the early years led to continual organisational difficulties and, in particular, problems in training a sufficient number of reliable examiners.

The high failure rate suggests the need to develop a lower-level examination so that weaker candidates can obtain recognition of their attainments. Some lower-level examinations do exist, but they do not attract great numbers of candidates, either because enrolment is limited to certain groups or because of class attendance requirements.

Figure 3 EGA



Data by courtesy of the Basque Government

Evolution of language examinations

Initially, in 1974, the Royal Academy of the Basque Language created the so-called *D titulua* to meet the demand of Basque medium schools for a guaranteed level of competence in the language and cultural knowledge of teachers seeking posts in those schools. A more advanced level was created at about the same time. In 1976, state-owned language schools started to offer an examination at a similar level. A smaller private body active in Basque culture set up a third, broadly similar, examination aimed at Biscay dialect speakers.

In 1979, Basque became an official language in the BAC, alongside Spanish. This brought about a major change in the status and future possibilities of the language: Basque language loyalists obtained access to substantial public funding for the furtherance of their aims. Shortly afterwards, at the end of 1981, in the BAC the Academy decided to pass on responsibility for examining, which it regarded as peripheral to its corpus planning activities, to the new regional government, which set up EGA as the successor of *D titulua*. Similarly, in 1986 the Academy also handed responsibility for examining in Navarre to the Government of Navarre, once the status of Basque had been clarified. It retains responsibility for examining in the French Basque Country in the absence of any official body with responsibility for the defence or promotion of the language.

In 1989 specific Basque-language requirements were introduced for certain BAC public administration posts, while in 1993 similar profiles were introduced for state-sector teachers, replacing the previous system primarily based on EGA and dating from 1983. The existence of a number of certificates, many purporting to certify a level similar to that of EGA, led to a period of confusion for candidates, who might have ended up having to sit more than one examination depending on where they wanted to use the resultant qualification. As the major examination, EGA led efforts between 1982 and 1998 to simplify the system by promoting the establishment of a series of administrative equivalences between certificates throughout the Basque Country, not just in the BAC. Since then it has focused attention on improving the quality of its own examination.

Forces shaping examinations

In addition to all the factors present in shaping majority language examinations, two extra sets of forces need to be mentioned in relation to Basque: the limitations due to Basque being a minority language and the constraints due to examinations being a conscious tool of language policy.

Limitations of a less-used language

First of all, there are the intralinguistic limitations: Basque is a language still undergoing development of its written standard. Additionally, like many less-used languages, Basque is not used (or as widely used) in as many domains or for as many functions as are larger mainstream languages. Secondly, there are the extralinguistic limitations: the market in Basque language teaching and testing is, by the standards of the major European languages, very limited indeed. Market size has obvious consequences for the possibilities of carrying out relevant research. Lack of a tradition in language teaching has moreover meant that many teachers have had access to only limited training. Much of the information on learning, teaching and testing now available is in English, of which Basques involved professionally in the field often have little or no command.

Examinations as a conscious language policy tool

The other contrast with mainstream languages has to do with the conscious use of examinations for language-planning goals in pursuit of the continued survival of the minority language. This has brought direct governmental control, with its advantages and disadvantages. Commercial objectives are not particularly valued; the examination system in itself is marginal to the main governmental concerns and is thus not prioritised (though the uses to which results are put are much discussed); the process of testing is excessively divorced from academic life.

A second aspect of this function of examinations as a language-planning tool is related to the choice of variant for examination. In major European languages the choice has often long been established; in languages such as Basque that are still undergoing standardisation, the choice is not always so straightforward. Standardisation plans laid before the Spanish Civil War did not come to fruition until 1968, when the standard written language, or *batua*, based on the geographically central dialects, was brought into being by the Academy. This choice of written standard, still the subject of ongoing decision-making, was nevertheless a source of conflict: one dialect group in the BAC – that speaking the westerly, Biscay dialect – lost out in relative terms, as linguistic distance from the standard is greater than that of other BAC dialects. The standard has been adopted in EGA and most other examinations, though most in the BAC also accept some use of the Biscay dialect, both in oral and written work.

Finally, examination setters have explored specific language-planning washback effects in addition to the more pedagogical goals habitually pursued: EGA has tried to influence classroom teaching by the weighting (50%) given to oral skills, the introduction of authentic texts for listening comprehension and the publication of past papers. On the planning washback side, examinations require standard spelling for all dialects and on the whole, at the written level, they give prominence to the standard over the dialects.

Examination content

With the teaching of Basque being unofficial and with no formal training for teachers available, it is hardly surprising that knowledge and practice of language-teaching methods among teachers of Basque under the Franco regime lagged behind those of teachers of the major European languages. Even at the end of the 1970s, grammar translation methodology seems to have been extremely common, with the first audio-lingual-based textbook appearing at the end of the decade. Communicative methodology made an even later appearance. Task-based learning is now popular, at least for the teaching of Basque to adults.

In line with the grammar-translation tradition, the written paper of the *D titulua* could contain more text in Spanish than in Basque; an oral component seems not always to have been included. The experimental 1982 EGA examination, established as standard the following year, represented a major departure: the cultural knowledge element was hived off and the examination focused entirely on the four skills. The oral examination was made a major permanent feature. After some initial experimentation with cloze tests and other exercises, which turned out to be unacceptable to examiners, the examination settled down to what is still its basic configuration:

- a written paper consisting of composition, reading comprehension, vocabulary questions and sentence transformations;
- oral skills tested by a listening comprehension test and an interview.

The composition was initially a traditional school-essay-type question, but candidates are now offered two alternative topics, each giving a title, a full-page collage of relevant brief quotes, pictures and/or statistical material, and an optional brief outline. Titles are usually chosen with the average age of candidates (young adults) in mind. Reading comprehension is examined by multiple-choice questions based on an authentic text. Vocabulary is tested by asking for appropriate substitutes for given words in a paragraph. Sentence transformations examining very varied aspects of the language are elicited by brief cues. The multiple-choice listening comprehension questions are based on one or two authentic texts, while the individual interview, usually carried out by two examiners, consists of two parts: one where the candidate gives an oral summary of an authentic written text and another where they provide an oral presentation on one of a number of topics offered. The candidate is allowed preparation time immediately before the interview for both activities, and may make notes for the presentation but not read a prepared text. A successful candidate has to pass all sections of both papers.

The only major change in format has been the subsequent introduction (1992) of a 100-item, eliminatory, computer-marked MCQ paper including both listening comprehension (based on authentic taped material) and written items based on a number of aspects of the language, to ensure that only those candidates with some possibility of passing the examination sit the more labour-intensive written paper and oral interview. Aspects of language tested include expressions taken from the spoken language, appropriateness, sayings, vocabulary, morphosyntactic items, knowledge of forms associated with the written standard, comprehension of the written language, and so on: the number of items of each type is predetermined. The introduction of this new paper has been monitored to ensure that overall pass rates have not been affected: in any case, the design of the scoring mechanism ensures that at least 50% of the candidates go on to the second, written, examination. Some teachers and examiners were concerned that the MCQ format would work to the detriment of native speakers unfamiliar with it and to the advantage of second-language learners who would have met it in class work. This, however, has proved not to be the case: statistics show that the performance of both groups continues to be similar to their performance before the introduction of the new paper. Owing to the large numbers of candidates sitting this paper on a single day, up to eight variants are produced to hinder copying in crowded examination halls. Where candidates appeal against the initial result, their examination is corrected again, this time by hand: the optical reader system, however, has proved to be extremely reliable.

Let us briefly review test construction: responsibility for setting papers according to the design established by the Department of Education, in consultation with the academic board, and published in the Basque Government Official Gazette lies with the part-time academic board directorate, consisting at present of the head examiner, the chief examiners of the three BAC provinces, plus two representatives of HABE, the government body responsible for overseeing the teaching of Basque to adults. Sheer volume of work means that proposals for specific papers or parts of papers are often prepared in the first instance by one or other of the examiners – at present about 120, generally practising, teachers – according to detailed written specifications provided by the directorate. This material is subsequently examined by the board directorate. Given the very broad range of language competence among present-day Basques, the authentic texts used are scrutinised to ensure that they represent a good quality of speaking or writing. Written texts are also edited to reflect present-day standard spelling and to remove excessively local or obscure linguistic features. Final decisions on specific points are normally reached unanimously by the six-person directorate.

No formal pre-testing is carried out, for security reasons: with such a small, demographically concentrated target population it has always been considered excessively risky. Quality control is basically through expert analysis by members of the board and its directorate. The administrative head of the organisation, present at directorate meetings, also contributes, particularly by actually sitting the MCQ test in advance to detect problem questions. *A posteriori* analysis of candidate performance on the test has helped in some cases to point to types of question that generally function well. No statistical analysis is carried out on other papers, so heavy reliance is placed on examiners.

Specific training is provided annually to examiners old and new to reduce divergence, with all examiners marking a number of practice papers and examiner performance being compared. Many written papers are in any case double-marked: a percentage of each examiner's papers and even all of an examiner's papers where deemed necessary, either because the examiner is new or because of doubts about his/her ability to mark appropriately. Where divergence is noticeable the two markers are expected to meet to discuss the paper and try to reach agreement; if they fail to do so, the chief provincial examiner intervenes. An appeals procedure ensures that written papers are further re-marked where necessary. Oral examinations can be taped if the candidate so requires, permitting a revision of the oral examination; on the whole, however, candidates do not request this option.

Finally, further measures have been taken over the years to avoid potential conflicts of interest among examiners in connection with the schools where they hold full-time jobs, or because they have relatives sitting the examination.

Issues pending

From this overview it is evident that much work remains to be done to bring EGA up to European standards. The examination body has adopted the ALTE Code of Practice, though this has brought only minor change as previous practice was largely along similar lines and the body has subsequently joined ALTE itself; the main focus of concern is now the academic aspect of the examination. The intention is to adapt it fully to the Common European Framework, which will no doubt mean changes in exercises and marking systems.

More statistical analysis of items would be desirable, though pre-testing in such a small society is difficult. Creation of an item-bank would also be desirable, but this runs against the tradition of rapid publication of past papers for pedagogical purposes and, increasingly, against the demands of open government, according to which some maintain that candidates have the right to see their paper after the examination and to copy any part they wish.

Greater academic back-up would undoubtedly be an advantage, but appropriate expertise through the medium of the Basque language is not easily obtainable; expertise from major language experts, while welcome, tends to fail to take specific local problems into account and needs adaptation.

17 Measuring and evaluating competence in Italian as a foreign language

Giuliana Grego Bolli
Università per Stranieri di Perugia, Italy

Aims

This paper aims to give an overview of the assessment of abilities or competencies in Italian as a foreign language, presenting the history and the current situation of certification in Italy and stressing the importance of the certification experience in order to promote a different 'culture' of assessment and evaluation in the Italian context.

The CELI examinations and qualifications structure will be also described.

The philosophical background

The philosophical background can be used as a reference point in order to explain how assessment and evaluation of competencies are perceived in the Italian educational system.

Empiricism (Hume, Locke), that is, the basis for empirical research on human behaviours, has never belonged to the Italian culture and way of thinking. Italian culture, also from an educational perspective, established its roots in Neo-Platonism and Cartesianism (Descartes), according to which 'knowledge' is an innate faculty and as such is equally distributed among human beings and does not need to be proved.

Besides that, the *Cogito, ergo sum* by Descartes is one of the strongest affirmations of Subjectivism, which is also a relevant feature of Italian culture and thought.

The empiricist perspective, on the contrary, is strongly opposed to any kind of innate principle, as well as to the consequent intellectual dogmatism, because it considers that Innatism is simply an *opinion* that cannot be justified, and it affirms that only through concrete experiences can human beings reach 'the knowledge' that is consequently restricted to the world of experience.

These basic considerations affect the method of investigation of human

sciences based on an inductive process from the perspective of Empiricism, and on a deductive one, according to the theories of Rationalism.

Nevertheless, educationalists have done a lot of work in this particular field and approaches based on empiricist theories have been introduced in the educational sciences in Italy since the beginning of the last century and, particularly after the Second World War. Nowadays the scenario seems to be starting to change within the Italian educational system as well.

The area of Linguistics

Italian Linguistics has a strong philological and theoretical tradition. Applied Linguistics is not yet a subject in its own right apparently, and it is not yet considered to be an academic discipline by theoretical linguists in Italy, perhaps because of its attempt to mediate between theoretical and practical applications or because of its problem-solving character. Nevertheless, Applied Linguistics is nowadays included in the area of linguistic disciplines within the Italian university system, but assessment and evaluation of L2 knowledge are not yet a matter of specific and systematic research and investigation.

The important contribution of a branch of the educational sciences, the so-called *Pedagogia Sperimentale*, together with the expansion of the certification experience all over Europe and the subsequent introduction of the concepts of scaling, of standardised tests and of defined assessment criteria, certainly influenced a new approach to the assessment and evaluation of competencies or abilities in Italian as a Foreign Language.

The beginning of the certification experience in Italy, the involvement of the University for Foreigners of Perugia with ALTE (Association of Language Testers in Europe) and with the Council of Europe, the *Progetto Lingue 2000* in Italy, as well as the recent publication of the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* by the Council of Europe could also make an important contribution to the introduction of a new perspective in the field of assessment and evaluation of language knowledge.

The certification experience

Italy has a quite recent tradition in the field of examinations and qualifications in Italian as a Foreign Language compared to much more widely spoken languages such as, for instance, English. The reason for that can be traced to the situation that I have outlined above, and also to the fact that Italian has been a language less analysed and described than other European languages, at least until the end of the 1980s.

The first certification experience at the University for Foreigners of Perugia started in 1987. The two levels of linguistic competence certified at that time represented an advanced phase in the learning process.

Why did we start, and why with two 'advanced' levels?

First of all there was a considerably large request for qualifications that could demonstrate knowledge of Italian as a foreign language; secondly, important research on the motivations for studying Italian was published in 1987 (Baldelli 1987). The research revealed that most people studying Italian as a foreign language were students or cultured people who were interested in learning Italian for study or cultural reasons; those people needed to know the language in order to handle most of the communicative situations that could occur in an academic or professional setting.

According to the above-mentioned research, the reasons for studying Italian all over the world were essentially four: culture, tourism, emigration, economic strength (Baldelli 1987).

With regard to the above-mentioned assertion, R. Simone (1989) pointed out some important considerations that to a certain extent can be considered still valid:

> il destino dell'italiano è, ancora una volta, quello di una grande cultura, l'attrazione che esercita è essenzialmente dovuta all'attrazione che esrcita la sua cultura e il mondo che essa esprime. La sua diffusione dipende vitalmente dalla diffusione della cultura italiana all'estero o, se preferite, l'italiano si diffonderà nello stesso modo e con lo stesso ritmo con cui accrediteremo l'italia e la sua reputazione. (1989: 109)

Nevertheless, during the last decade a unique combination of social, economic and political factors in Europe have led to a considerable change in the motivation for studying the most important European languages, including Italian.

People were interested not only in reaching high levels of proficiency, but also in lower levels that could guarantee a fairly active participation in an informal conversation and an exchange of information with the inhabitants.

The current situation

In 1993 a new phase began. An agreement was signed between the Universities for Foreigners of Perugia and Siena and the Ministry of Foreign Affairs, according to which the Ministry recognised our certifications and designated the Italian Institutes of Culture abroad as the official seats of the examinations produced by the two Universities. Afterwards the same agreement was also signed by the Third University of Rome. Nowadays the examinations and certificates produced by the 'Dante Alighieri' are also recognised by the Ministry. Currently in Italy there are four certification systems that are officially recognised; of course it is not easy to compare them in terms of levels, competencies assessed and approaches adopted. To date, in fact, no comparative studies have been conducted that are capable of demonstrating any possible parallels between the four certification systems.

The number of certificates in Italian awarded by public, academic and private institutions operating in Italy or abroad is nowadays increasing because of a general demand for certificates attesting to a knowledge of the language, but we should carefully consider that certificates can have a strong influence and a powerful impact, both on society and on the education system and hence, specific, professional and technical competencies are needed during the development and production process of the exams.

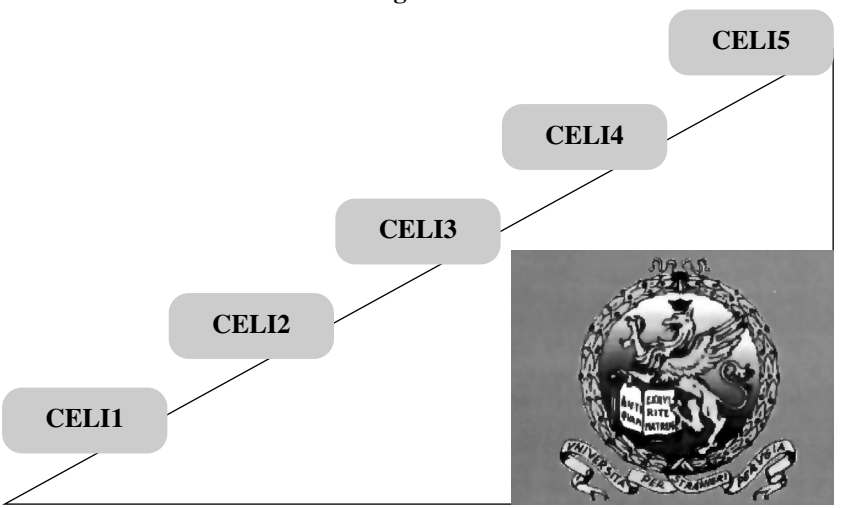
To certificate the knowledge of a language in terms of ability to use it is in itself a risky and difficult challenge; it implies a process of inference and generalisation that involves a rigorous theorisation and a consequently thorough process of elaboration, trial, analysis and production of materials, which is very demanding in terms of competencies and of economic and human resources required. All these considerations need to be taken into account before starting any certification project.

The CELI examinations

CELI is the acronym for Certificates of Italian Language. They are produced and distributed by the University for Foreigners of Perugia. The CELI system started operating in June 1993 all over the world. The CELI examinations are aimed at assessing the knowledge of general Italian and are not intended for specific purposes; they are administered twice a year (in June and November) both in Italy and abroad.

The scale adopted is a five-level proficiency scale starting from a fairly 'low' level, CELI1, and reaching a fairly 'high' level, CELI5.

Figure 1



Each examination encompasses different components, grouped according to the four basic abilities: reading, writing, listening and speaking. Nevertheless, skills are not assessed separately but in an integrated way, as they occur in the reality of communication. This integration of abilities, which needs to be constantly monitored, could be a cause of major problems in the evaluation phase.

Starting from level 3, a specific component was introduced in order to assess the knowledge of grammatical and lexical elements.

The results are reported in terms of a grading scale A-E, with C as the Pass and D and E as negative grades.

So far around 28,000 candidates have taken the CELI examinations all over the world. The most often awarded level in the CELI proficiency scale is level 3, that is to say an intermediate level, and since 1998 the great majority of CELI candidates have reached a level between levels 1 and 3, as is clearly shown in the table below (the data reported in the diagram stop at the end of 2000 and do not include the last session of June 2001).

Since June 2000 a complementary examination system called CIC (Certificati di Italiano Commerciale) has been set up in order to assess the knowledge of Business Italian. The system includes tests divided into the four basic language skills and it comprises two levels: CIC Intermediate (corresponding to level 2 on the CELI scale) and CIC Advanced (corresponding to level 4 on the CELI scale). The CIC exams are administered only once a year, at the end of June.

Table 1 Examination distribution from 1993 to date by levels

Anno	CELI1	CELI2	CELI3	CELI4	CELI5	CIC	Total
1993	27	40	183	28	79	86	443
1994	73	174	574	226	189	121	1357
1995	175	397	910	330	276	108	2196
1996	175	544	1086	440	339	120	2704
1997	229	676	1687	478	422	103	3595
1998	259	1033	2066	664	465	95	4582
1999	324	1044	2271	751	590	24	5004
2000	387	1148	2585	839	544	42	5545
Totali	1649	5056	11362	3756	2904	699	25426

The number of CELI candidates has increased constantly during the last eight years and the great majority of our candidates, as shown in the following table, clearly prefer the June session to the November one.

Table 2 Number of candidates participating in each session from 1993 to date

Year	1993	1994	1995	1996	1997	1998	1999	2000
June	228	922	1332	1660	2390	2735	3065	3430
November	215	435	864	1044	1205	1847	1939	2115
Total	443	1357	2196	2704	3595	4582	5004	5545

There are also some considerations relating to the age and the gender of our candidate population. They are mostly quite young, between 18 and 25 years old, and 80% are women:

Table 3 Subdivision of candidates by age from 1993 to date

Age	0–18	18–25	25–35	35–45	>45
1993	15	110	197	73	48
1994	140	662	364	115	76
1995	284	991	585	194	142
1996	370	1319	694	187	134
1997	518	1901	838	198	140
1998	639	2394	1042	276	231
1999	594	2664	1266	290	190
2000	630	2876	1384	300	355

Table 4 Subdivision of candidates by gender from 1993 to date

Year	1993	1994	1995	1996	1997	1998	1999	2000
Men	79	302	449	546	736	912	953	1082
Women	364	1055	1747	2158	2859	3670	4051	4463
Total	443	1357	2196	2704	3595	4582	5004	5545

The data reported in the above two tables are important not merely from a statistical or theoretical point of view; they need, in fact, to be taken into account in the selection of the materials to be used in the exams.

How did we develop the CELI proficiency scale?

Developing a proficiency scale means, in practice, placing activities and competencies at different levels. The points on the scale at which particular activities and competencies should be situated are generally determined objectively, with reference to a measurement theory. At the very beginning the CELI proficiency scale was worked out in a purely intuitive way according to quite subjective criteria, on the basis of the experience we had as teachers of Italian as a Second Language and according to descriptive categories reported in grammars of Italian and in lexical lists. Afterwards, the effectiveness of the categories described was checked with teachers of Italian abroad by means of questionnaires, seminars and refresher courses, etc.: a sort of qualitative, descriptive approach that needs, however, to be empirically proved, rather than a quantitative approach based on measurement theories. The University of Foreigners of Perugia, according to the ALTE Code of Practice, is in the process of starting to implement a measurement model such as the Rasch model in order to validate the CELI proficiency scale.

The CELI examinations: number of levels

First of all some basic assumptions: levels are intended to be learning objectives; the decision about the number of levels to refer to depends mostly on the context in which they would be used. In a certification perspective they should be adequate to cover the range of competencies and abilities required for the target population, but *'should not exceed the number of levels between which people are capable of making reasonably consistent distinctions'* (*Common European Framework* 2001: 21). The levels described by the Council of Europe were from the beginning a stable point of reference for the definition of the CELI proficiency scale and of the ALTE scale as well, and the general acceptance of the *Common European Framework* proficiency scale is nowadays a confirmation of the correctness of the decision taken jointly by the University of Foreigners and by ALTE in the early 1990s.

The level descriptors (G. Grego Bolli and M. G. Spiti 2000) have also been formulated in a transparent way in order to be clearly understandable by non-specialists as well as, for instance, by students. They were intended to describe concrete tasks and the degrees of skill used in performing those tasks. We tried to keep the descriptors as concise as possible because teachers and users in general prefer short descriptions.

Level descriptions

ALTE level 1 CELI1 Waystage FA2	ALTE level 2 CELI2 Threshold FB1	ALTE level 3 CELI3 Vantage FB2	ALTE level 4 CELI4 Effective proficiency FC1	ALTE level 5 CELI5 Mastery FC2
Can understand simple sentences and expressions used frequently in areas of immediate need. Can exchange basic information on familiar and routine matters of a concrete type.	Can deal with most situations related to tourism and travelling. Can produce simple connected texts on familiar topics or on subjects of personal and everyday interest.	Can understand the main points of concrete or abstract texts. Can interact with native speakers with an acceptable degree of fluency and spontaneity without particular strain. Can explain his/her point of view and express opinions.	Can understand longer and more complex texts, making inferences. Can express him/herself fluently, spontaneously and effectively. Can produce well connected and structured texts showing control of cohesive devices.	Can understand virtually every text heard or read. Can reconstruct arguments and events in a coherent and cohesive presentation. Can express him/herself spontaneously, very fluently and precisely. Can understand finer shades of meaning.

The correspondence between the CELI proficiency scale and the ALTE scale

There is a clear correspondence (in number of levels, tasks and performance required, etc.) between the ALTE five-level system (*ALTE Handbook* 1998) and the CELI five-level system: the two developed together at the beginning of the 1990s. It will take much more time and effort to demonstrate that correspondence empirically. Some years ago a long-term, on-going project called '*Can-do*' was undertaken by the ALTE members in order to develop and validate a set of performance-related scales across the range of languages represented by ALTE. So far, data collection has been based on self-report, the 'Can-do' scale being presented as a set of linked questionnaires. Following that phase, attention has turned to establishing the link between the 'Can-do' scales and performance in ALTE examinations. The University for Foreigners of Perugia, in collaboration with the EFL Unit of the University of Cambridge Local Examinations Syndicate, started looking in 2000 at the performance in CELI examinations and collecting data to link 'Can-do' self-rating to grades achieved in CELI exams at different levels. The preliminary data collected and analysed showed quite a good relationship; however more data need to be collected to confirm this preliminary result.

It would seem possible (and useful for the description of language proficiency) to align the ALTE framework to the Council of Europe framework of levels; in fact the descriptions provided by the ALTE framework are in many ways complementary to those of the CEF and both systems have used similar statistical approaches. A study aimed at relating the two frameworks has been undertaken by the EFL Unit of the University of Cambridge Local Examinations Syndicate. The preliminary results of this study have been published in Appendix D of the *Common European Framework* (pp. 244–50).

The CELI elaboration and production process

A special unit has been established within the University to deal with all the aspects involved in the elaboration and administration of the exams. First of all we agreed on starting from a model for the definition of the abilities to be tested, and we chose the model elaborated by Weir in 1993. The model defines each ability in terms of: *operations, conditions and levels* (Weir 1993).

According to the model, a team of expert teachers is then delegated for the selection of materials and for the test item-writing. They participate in the editing process, which is led by team leaders. Materials written for Reading, Structural Competence and Listening are pre-tested on foreign students at the University.

Item writers are responsible for selecting materials from a wide range of authentic sources: newspapers and magazines, brochures, forms, contracts, advertisements, radio items, books, etc.

Materials are selected on the basis of:

- Skills to be assessed; activities to be carried out
- Linguistic complexity (relating to the learning objectives for each level)
- Text type (relating to the degree of familiarity of the candidate population with the text genre and the domain)
- Subject matter and cultural consideration (in order not to cause negative reactions in the candidates)
- Reference to lexical lists
- Relevance to the learners' personal interests and socio-cultural background
- Item types to be used

Item types

Item types have to be selected first of all in accordance with a linguistic theory – the definition and description of the construct that we are going to assess through the performance of candidates in the exams – but also taking into account practical considerations such as the number of candidates, which is usually quite high in a certification context.

Besides that, language competence, such as the ability to communicate in a L2, is such a complex multidimensional phenomenon that various item types are needed in order to assess different skills and sub-skills.

We tried to keep a good balance between objective tasks (multiple-choice, multiple-choice gap-filling, gap-filling, matching, editing, transfer information, sentence transformation) and subjective tasks (composition, guided composition, essay, summary, report, open-ended question, guided conversation). In fact, there are students who are quite familiar with objective items and know how to observe tight time limitations, but there are also students who are more familiar with subjective items such as compositions or essays; these students need time and, possibly, rough copies for their writing. Of course all these different backgrounds should be taken into account in order not to advantage or disadvantage candidates.

Clearly it does not make any sense to say *a priori* that objective items are better than subjective items, or vice versa; there are a number of different considerations at the basis of the item-type selection and any choice can be, at least, justified.

Objective items are easy to correct and evaluate, but are really time-consuming to prepare and often not as 'efficient' and significant, at least without a careful pre-testing phase and an adequate items analysis. Subjective items, on the other hand, tend to elicit a better performance and are relatively easy to prepare, but they are extremely difficult to evaluate, and they require specific criteria and grading scales to refer to and special preparation and training for examiners.

As an educational institution we are convinced that one of our major commitments and tasks is the preparation of professional examiners, and we are currently undertaking this, which is an acceptable reason for using quite a few subjective items.

The marking process

All the candidates' papers are sent back to Perugia and marked at the Certification Unit. The examiners are either practising teachers or trained external personnel working under the guidance of team leaders; objective tests are marked on the basis of mark schemes. We are going to set up computerised methods of marking, involving optical mark readers.

The writing components are marked on the basis of specific rating scales and sample scripts.

Rating scales for writing

Different kinds of task generally make up the writing component. We use more contextual tasks and guided compositions than free compositions in order not to have different productions, because the latter are more difficult to mark according to the standards.

The rating scales have been formulated defining descriptors for four assessment criteria: vocabulary control, grammatical accuracy, socio-linguistic appropriateness and coherence.

Vocabulary control	**Grammatical accuracy**	**Socio-linguistic appropriateness**	**Coherence**
Command and adequacy of the lexical repertoire (vocabulary) in order to be able to carry out an assigned task. Orthography	Knowledge and use of morphosyntactic processes (formal word modifications) and of connecting mechanisms.	Capacity to use the language in context, respecting the aim of expressive ability and in connection with the situation and/or argument treated.	Capacity to produce written texts that indicate thematic continuity and effective expressive ability.
Padronanza ed adeguatezza del repertorio lessicale (vocabolario) per portare a termine il compito assegnato. Ortografia	Conoscenza ed uso di processi morfologici (modificazioni formali delle parole) e di meccanismi di collegamento.	Capacità d'uso dell lingua in contesto, nel rispetto delle finalità comunicative e delle relazioni di ruolo richieste da una situazione e/o dall'argomento trattato	Capacità di produrre testi scritti che dimostrino continuità tematica ed efficacia comunicativa

Standards for each criterion have been defined by a sliding scale of 1 to 5, where 5 is 'very good' and 1 is 'very weak' (Grego Bolli and Spiti 1992).

We have not taken into account criteria such as adequacy of input or relevance of content, because we were convinced that these were not primarily linguistic criteria and do not reflect the essential linguistic competence of a writer. Furthermore, in the light of our ten-year experience, we detected a need to be more precise and transparent in defining how a non-relevant (or not entirely relevant) content will affect the score and how we should score a script where not all the parts of the tasks were addressed or fully addressed.

Rating scales for speaking

The speaking component comprises a face-to-face interview based on material (text, pictures, etc.) used as a prompt. All the materials are sent from Perugia, according to the number of candidates, in order to have as uniform a production as possible.

The performance of each candidate is marked locally by two trained examiners (generally mother tongue teachers of Italian), following grading scales. Generally the two examiners also act as interlocutors. Marks are recorded on registers and sent to Perugia. The grading scales have been

formulated defining descriptors for four assessment criteria: vocabulary control, grammatical accuracy, socio-linguistic appropriateness, pronunciation:

Vocabulary control	**Grammatical accuracy**	**Socio-linguistic appropriateness**	**Pronunciation**
Has a good command and adequacy of the lexical repertoire (vocabulary) in order to be able to carry out an assigned task	Knowledge of the morphosyntactic mechanisms necessary for the functional varieties of the spoken language	Capacity to express oneself in an adequate manner according to the situation (argument, role of the interlocutors) and to conduct a conversation with respect to the typical conventions of spoken language	Capacity to express oneself in a comprehensible and acceptable manner. Capacity to point out the type of enunciation (affirmative, interrogatory) and its pragmatic value (joke, sarcasm, etc.)
Padronanza ed adeguatezza del repertorio lessicale (vocabolario) per eseguire il compito assegnato	Conoscenza dei meccanismi morfosintattici necessari alle varietà funzionali del parlato	Capacità di esprimersi in modo adeguato alla situazione (argomento, ruolo degli interlocutori) e di condurre una conversazione nel rispetto delle convenzioni tipiche del parlato	Capacità di sapersi esprimere in modo comprensibile ed accettabile. Capacità di segnalare il tipo di enunciato (affermativo, interrogativo) e il suo valore pragmatico (scherzo, ironia, ecc)

Standards for each criterion are defined by a sliding scale of 1 to 5: where 5 is 'very good' and 1 is 'very weak' (Grego Bolli and Spiti 1992).

We are considering a number of changes in relation to the construction and evaluation of the speaking component.

First of all, it has been proven that, during a face-to-face interview, candidates generally tend to perform the same functions (expressing opinions, giving information, describing, etc.): there is almost no negotiation involved because of the exam situation and the dominant position of the examiner.

Secondly, it has been proven that an examiner who also acts as the interlocutor tends to evaluate his or her performance together with or instead of the performance of the candidate.

We are considering, therefore, the possibility of having an interview with two candidates on one side and an interlocutor plus one examiner on the other. The examiner should not be directly involved in the interview, but should only mark the candidates' performance.

These changes would certainly imply the production of different tasks for the speaking component of the exam and new preparation and training for both the interlocutors and the examiners.

Training for examiners

For written papers, marking schemes and sample scripts are discussed by team leaders and examiners in order to standardise marking as much as possible.

For the speaking component, training involves examiners coming to Perugia for a short course, and team leaders or external consultants giving seminars in Italy and abroad which use videos to ensure an acceptable level of standardisation of assessment.

Conclusions

We have provided a brief history of language certification in Italy and of the certification experience of the University for Foreigners in Perugia. Of course the system formulated in Perugia can be improved, and it will need some changes in the near future, but nonetheless the certification experience has made an important contribution both to the knowledge of Italian all over the world and to the introduction of a new perspective in the field of assessment and evaluation of competencies in Italian as a Foreign Language.

References

ALTE Handbook of Language Examinations and Examination Systems.

Baldelli, I. (ed.) 1987. *La Lingua Italiana nel Mondo. Indagini sulle Motivazioni allo Studio dell'Italiano*. Roma: Istituto della Enciclopedia Italiana.

Grego Bolli, G. and M. G. Spiti. 1992. *Verifica del Grado di Conoscenza dell'Italiano in una Prospettiva di Certificazione. Riflessioni, Proposte, Esperienze, Progetti*. Perugia: Edizioni Guerra.

Grego Bolli, G. and M. G. Spiti. 2000. *La Verifica delle Competenze Linguistiche. Misurare e Valutare nella Certificazione CELI*. Perugia: Edizioni Guerra.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Simone, R. 1989. *Il Destino Internazionale dell'Italiano*. In 'Italiano ed Oltre', 4: 105–109.

Weir, C. J. 1993. *Understanding and Developing Language Tests*. Hemel Hempstead: Prentice Hall.