Language Testing Matters

Investigating the wider social and educational impact of assessment

Proceedings of the ALTE Cambridge Conference, April 2008

Also in this series:

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 equation modeling approach
James E. 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
Edited by 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
Edited by Cyril J. Weir, Yang Huizhong, Jin Yan

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 J. Weir and Michael Milanovic

A Modular Approach to Testing English Language Skills: The development of the Certificates in English Language Skills (CELS) examination
Roger Hawkey

Issues in Testing Business English: The revision of the Cambridge Business English Certificates
Barry O'Sullivan

European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference July 2001
Edited by Cyril J. Weir and Michael Milanovic

IELTS Collected Papers: Research in speaking and writing assessment
Edited by Lynda Taylor and Peter Falvey

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

Changing Language Teaching through Language Testing: A washback study
Liying Cheng

The Impact of High-stakes Examinations on Classroom Teaching: A case study using insights from testing and innovation theory
Dianne Wall

Assessing Academic English: Testing English proficiency 1950–1989 – the IELTS solution
Alan Davies

Impact Theory and Practice: Studies of the IELTS test and *Progetto Lingue 2000*
Roger Hawkey

IELTS Washback in Context: Preparation for academic writing in higher education
Anthony Green

Examining Writing: Research and practice in assessing second language writing
Stuart D. Shaw and Cyril J. Weir

Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity. Proceedings of the ALTE Berlin Conference, May 2005
Edited by Lynda Taylor and Cyril J. Weir

Examining FCE and CAE: Key issues and recurring themes in developing the First Certificate in English and Certificate in Advanced English exams
Roger Hawkey

Language Testing Matters

Investigating the wider social and educational impact of assessment

Proceedings of the ALTE Cambridge Conference, April 2008

Edited by

Lynda Taylor
Consultant
University of Cambridge ESOL Examinations

and

Cyril J Weir
Powdrill Professor in English Language Acquisition
University of Bedfordshire

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

www.cambridge.org
Information on this title: www.cambridge.org/9780521163910

First published 2009

Printed in the United Kingdom at the University Press, Cambridge

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

Library of Congress Cataloging-in-Publication data
ALTE Conference (3rd : 2008 : Cambridge, England)
Language testing matters : investigating the wider social and
educational impact of assessment : proceedings of the ALTE Cambridge
Conference, April 2008 / edited by Lynda Taylor and Cyril J Weir
p. cm – (Studies in language testing : v. 31)
ISBN 978–0–521–16391–0
1. Second language acquisition–Ability testing–Congresses. 2.
Second language acquisition–Social aspects–Congresses. 3. Language and
languages–Ability testing–Congresses. I. Taylor, Lynda. II. Weir,
Cyril J. III. Title. IV. Series.
P118.75.A48 2008
401'.93--dc22

ISBN 978-0-521-16391-0 Paperback

Contents

Acknowledgements

We would like to express our thanks to all the volume contributors for developing and writing up their original presentations given at the ALTE Cambridge Conference in April 2008, and for their willingness to make subsequent revisions in line with our editorial suggestions.

The volume could not have reached publication without the professional, technological and administrative assistance of various staff members based at Cambridge ESOL including: Martin Nuttall in the ALTE Secretariat; Carrie Warren in the Research and Validation Group; and Sally Downes in the Stakeholder Relations and Legal Affairs Group. We are grateful to all of them for their support throughout the production process.

Finally, the publishers are grateful to the copyright holders for permission to use the copyright material reproduced in this book. Blackwell Publishing Ltd for Table 2.4 from *Language Testing: The Social Dimension* by T McNamara and C Roever, 2006; UK Home Office for a table from the *Accession Monitoring Report, 2008*; Office for National Statistics for a table from the *2001 Census*; Qualifications and Curriculum Development Agency for sample material from the *Adult Literacy Test, 2004*; Department for Business, Innovation and Skills for material from *Skills for Life* and *Pathways to Proficiency*, 2003; Learning and Skills Improvement Service for extracts from the *ESOL and Literacy Core Curriculum*; Oxford University Press for a table from 'Everyday creativity in language: textuality, contextuality, and critique' by J Maybin and J Swann, 2007, previously published in *Applied Linguistics* 28 (4); National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles, California, for the table previously published on page 6 of 'Accommodation strategies for English language learners on large scale assessments: Student characteristics and other considerations', (CSE tech. Report no. 448) by F A Butler and R Stevens, 1997; SAGE Publications Ltd for a figure in 'Working for washback: a review of the washback concept in language testing' by K Bailey, 1996, previously published in *Language Testing* 13; ABC-CLIO, LLC for Figures 5–7 from *Input Processing and Grammar Instruction in Second Language Acquisition* by W VanPatten, 1996; Hans Kåre Fløø for permission to use Figure 1 on page 345 and Johannes Hjellbreke for permission to use Table 1 on page 346.

Series Editors' note

The 3rd International Conference of the Association of Language Testers in Europe (ALTE) was held in Cambridge in April 2008, hosted by University of Cambridge ESOL Examinations, a non-teaching department of the world-famous Cambridge University. This third conference built upon the success of two previous ALTE Conferences: the first held in Barcelona in July 2001, hosted by the Generalitat de Catalunya, on the theme of 'European Language Testing in a Global Context' to celebrate the European Year of Languages; the second in Berlin in May 2005, hosted by the Goethe-Institut, on the theme of 'Language Assessment in a Multilingual Context' to support the 50th Anniversary of the European Cultural Convention. Edited proceedings from both events were published as Volumes 18 and 27 in the now well-established and highly regarded *Studies in Language Testing* series.

The theme of ALTE's 3rd International Conference – 'The Social and Educational Impact of Assessment' – was particularly topical in 2008, given the level of public debate around the use of language tests in the context of migration and citizenship, higher education and employment. A key challenge for us all is to ensure that the tests we provide are not just accurate, reliable, transparent and fair, but that they are explicitly designed to have as positive an impact on society as possible.

The Cambridge conference formed part of the International Year of Languages 2008, announced by the United Nations General Assembly in 2007. As language issues are central to UNESCO's (United Nations Educational, Scientific and Cultural Organization) mandate in education, science, social and human sciences, culture, communication and information, the organisation was named the lead agency in promoting this special year. UNESCO invited governments, United Nations organisations, civil society organisations, educational institutions, professional associations and all other stakeholders to increase their own activities to promote and protect all languages, particularly endangered languages, in all individual and collective contexts. As a non-governmental organisation (NGO) with special consultative status with the Economic and Social Council of the United Nations, ALTE offered its 3rd international conference in support of the International Year of Languages.

ALTE 2008 was one of the largest language testing conferences ever to have taken place. Over 300 abstracts were presented for consideration by the panel from which over 150 papers were accepted. Well over 500 delegates

attended from more than 50 countries around the world representing over 300 organisations. It was also a multilingual event, with presentations in five of the many different languages represented at the conference. The interest ALTE 2008 generated demonstrates the growing importance of language assessment in today's world as well as the increasing number of experts who are playing ever more important roles in policy making and implementation.

The conference organisers were particularly pleased to offer a forum on Language Testing, Migration and Social Inclusion, held under the auspices of the Secretary General of the Council of Europe, Mr Terry Davis. The forum focused on the work of European bodies, including the European Union and the Council of Europe, in relation to the integration and social inclusion of migrants and an exploration of intercultural dialogue. This contributed to the work programme of the Language Policy Division of the Council of Europe and specifically its project on language policies and the integration of adult migrants. The conference also welcomed the European Commission Directorate General for Education and Culture to talk about the important work of the European Indicator of Language Competences and its context.

The ALTE Cambridge conference marked another important stage in ALTE's development since it was originally founded with eight members in 1990, primarily to work on common levels of proficiency and common standards for the language testing process. Over nearly two decades, the association has contributed to a wide range of important international projects, many of which were featured during this and previous conferences and which are described in the opening pages of the conference proceedings volumes. Membership of ALTE has grown to the present total of 31 members – including many of the world's leading assessment bodies – who between them represent the testing of 26 languages. Europe thrives on diversity and it is the need to respect and value this diversity while at the same time trying to find common ground that binds us together. The event in Cambridge was a great opportunity for all of us to participate in a conference that reflects the diversity of Europe and the importance we all place in languages, language learning, the certification of language competence and the significance of the Common European Framework of Reference for Languages in the development of plurilingualism and intercultural competences. But improving mutual understanding is equally important in the wider global context. The event in Cambridge was a gathering of assessment professionals focusing not only on professional matters in our field but also engaging positively in debate on language in a social, economic and political context. If our voices are to be heard then we need to participate positively. We also need the capacity to see things from a number of perspectives and in organising this conference, one of the main aims was to allow for the divergence of views, opinions and perceptions in order to help this process of building mutual understanding.

ALTE provides a forum where assessment professionals can work together effectively and there are two particular projects where ALTE members have been working together effectively in recent years. The first is in relation to the survey of language competences in Europe. The survey was first mooted in March 2002, as part of a European Council strategy to 'improve the mastery of basic skills, in particular by teaching at least two foreign languages from a very early age'. Invitations to tender were issued in mid 2007 and SurveyLang, a consortium made up largely of ALTE members and led by Cambridge ESOL, was finally confirmed as the successful bidder in February 2008. The survey will provide information on the general level of foreign language knowledge (in five languages: Italian, French, German, Spanish and English) of the pupils in 32 Member States and other participating countries. It will provide strategic information to policy makers, teachers and learners in all surveyed countries and it is anticipated that the collected data from the survey will help policy makers, teachers and practitioners to take decisions about how to improve foreign language teaching methods and thus the performance of pupils in foreign languages. This is an enormously challenging but also potentially extremely useful project.

ALTE members are also working increasingly with national ministries of education to help provide high quality language assessment. The Lingua 2000 project in Italy was a highly successful example of this work some years ago when the Italian government made great use of international language certification to help in the learning and teaching of languages in Italian schools. More recently, ALTE members – the Cervantes Institute, the Goethe-Institut and Cambridge ESOL – have been working with the French Ministry of Education to provide language testing materials to international standards and linked to the CEFR (Common European Framework of Reference for Languages).

While seeking opportunities for effective collaboration and while seeking conformity to standards of good practice, respecting and understanding our differences is a key aspect of ALTE's work. To advance these aims, ALTE has developed guidelines for the writing of test materials, ways of describing the content of examinations so that they can be compared more effectively. ALTE has also built a framework of examinations that allows users to see how the different exams relate to each other and, importantly, members of ALTE have defined a multilingual glossary of language testing terms (developed and published in 10 languages in the late 1990s and now available in numerous additional languages). For instance, the latest edition was published in Basque in 2007, clearly demonstrating the sustainability of ALTE's work. Much of this work has been supported by funds provided by the European Commission through its Lingua programme, and much of it has been done in collaboration with the Council of Europe which has played and continues to play such a significant role in language policy in Europe and now through the

CEFR far beyond. All of this work is available on the ALTE website or from individual members of ALTE.

ALTE published its first international code of practice for language testing in 1994 and much work on refining this concept and documenting principles of good practice took place in the 1990s. Between 2000 and 2002 ALTE set up a Code of Practice Working Group which developed a Quality Management System leading to a Quality Auditing System that was piloted in 2005 and 2006 and introduced in 2007. ALTE has now audited many of its members on at least one of the examinations they provide. As a consequence of these developments, membership of ALTE is now based on demonstrating, through the Auditing System, that an organisation's examinations do conform to internationally recognised systems in a transparent and open way.

ALTE is in the process of developing web forums in English, French, German and Spanish in the first instance and we anticipate that the number of language forums will increase over the next few years. Within these, members will have access to the ALTE network, training materials, publications and training courses offered by ALTE throughout Europe on a relatively frequent basis.

Most recently, ALTE has taken steps to widen participation in its activities by bringing in new categories of Institutional and Individual Affiliates, allowing a wider range of organisations and individuals to make a real contribution to the development of a truly international approach to language testing. One of ALTE's main aims is to share ideas and know-how. Events such as the 2008 ALTE Conference in Cambridge provide an ideal opportunity for language teaching and testing professionals from around the world to meet and to pool expertise, and to consider together how best to resolve some of the important challenges facing society today. Not surprisingly, plans are already well in hand for a 4th ALTE International Conference to be held in Krakow, Poland, in early July 2011.

A full listing of all the presentations given at the ALTE 2008 Conference can be found at the end of this volume. As will be apparent, the 20 conference papers presented here represent only a selection of the many excellent presentations made in Cambridge reflecting a wide range of topics and concerns; they provide a flavour of the key themes addressed at the conference. The Introduction to this volume by Lynda Taylor and Cyril J Weir helps to highlight and summarise for readers the various strands that resonated throughout the conference, and points to important implications for the language testing community.

Michael Milanovic
Cyril J Weir
April 2009

Introduction

Lynda Taylor and Cyril J Weir

A stated aim of the *Studies in Language Testing* (SiLT) series is to support work in the related fields of applied linguistics and language testing by highlighting recent theoretical and practical developments in language assessment and by reviewing the impact these are currently having on education and society. *Language Testing Matters* – the 31st title to appear in the SiLT series since 1995 – embodies that aspiration. The volume brings together a selection of 20 edited papers based on presentations given at the 3rd ALTE International Conference, held in Cambridge in April 2008 and which took as its theme 'The Social and Educational Impact of Language Assessment'. The papers explore the social and educational impact of language testing and assessment, grouped according to three core themes:

- new perspectives on testing for specific purposes
- insights on testing in language teaching and learning
- reflections on the impact of testing among stakeholder constituencies.

This volume is designed to broaden our horizons so that we better understand the social and educational impact of language testing and assessment, and the extent to which language testing matters to individuals, to groups, to organisations, and to society as a whole. Given this aspiration, we trust it will be a valuable resource not only for the language testing and assessment community but also for the wider world of public policy and social concern, nationally, regionally and internationally.

Section One of the volume considers some fresh perspectives on testing and assessment for specific purposes in particular contexts of use. Although the field of language testing for specific purposes has been well-established for many years, shifting demographics due to globalisation as well as changing requirements and expectations regarding language competencies in study, workplace and social contexts have all led to a role for language tests in new and sometimes unexpected domains. These include professional fields such as medicine and the law, business domains such as the aviation and construction industries, and, perhaps most controversially, the social dimension of migration and citizenship policy. Furthermore, in our post-9/11 world, the certification of high-level language skills continues to be a priority for agencies charged with responsibility for national and global security. As the international geo-political landscape fragments into ever more complex

and unpredictable patterns, so governments and agencies find themselves drawing on intelligence sources that come from every corner of the world and in hundreds of different languages, with the consequent need to train, assess and certificate the foreign language proficiency skills of their personnel.

The opening paper in this volume by **Rachel Brooks and Beth Mackey** considers the development of tests for Less Commonly Tested Languages (LCTLs) in the context of the United States Federal Government's language proficiency assessment programme, particularly for the testing of receptive skills and oral proficiency. They discuss how the US Government, with its long tradition of language testing linked to foreign intelligence and national security, has sought to adapt its traditional models of test development, validation, and tester training to provide solutions that satisfy professional standards as well as its own institutional requirements. The two papers that follow move us into the fields of business and the professions. **Philip Shawcross** explores the complex relationship between the assessment of English language proficiency and the world of air travel in light of recent United Nations efforts, through the International Civil Aviation Organization (ICAO), to establish a global Language Proficiency Requirement for the aviation industry. He reviews the impact of aviation language testing from safety, professional, social and economic standpoints, showing how compliance with the high-stakes and operationally focused requirements of an aviation English testing process represents a very significant challenge for language testers. **Margaret van Naerssen** takes us into the legal world of the courtroom where the increase in court appearances of non-native speakers (NNSs) is resulting in the growing importance of linguistics in legal cases and thus an increased role for language testers and their assessment procedures. She demonstrates how those involved in the use of language assessment in such contexts need to be well prepared to make informed decisions in the use of particular protocols, to evaluate the choices, findings and use of results by others, and to deal with the legal challenges that will invariably be raised.

The growing use of high-stakes language testing around the world has brought social concerns about cheating to the fore over recent years. Cheating on tests, especially on high-stakes tests, is generally acknowledged to have significant impact, potentially resulting in serious consequences for test takers, for test score users, and for society more broadly, though to date there has been relatively little empirical research reported in our field on this phenomenon. Two papers in this volume offer us rare and fascinating insights into the world of cheating on language tests. **Dayong Huang and Mark Garner** report the findings of a research study examining the prevalence of cheating on a high-stakes language proficiency examination in one specific higher education context; they also explore the factors that motivate test takers to cheat, the counter-measures taken, and the consequences that result from cheating. A second paper by **Rachel Brooks** also addresses the phenomenon of

cheating in tests, but this time from the perspective of a testing agency seeking to develop appropriate techniques for detecting cheating in a different type of high-stakes assessment – translation tests. She describes a research study that trialled authorship attribution techniques (more commonly used in academic research and in the courtroom) to determine whether such techniques can provide a replicable, valid procedure for detecting plagiarism on translation tests for Arabic and Urdu.

The final two papers in Section One focus on the social impact of language testing and assessment in the context of migration and citizenship, a public policy arena that has attracted a high level of attention and concern, and indeed criticism, from the language testing community in recent years. **Philida Schellekens** reflects upon what can be learned from the experience of the Skills for Life Strategy, launched in England and Wales in 2001 to improve the basic literacy and numeracy skills of the general population, both in terms of its expected and its unexpected consequences. One noticeable impact, for example, is that the achievement of migrants and refugees, who need English for social and work purposes, is being assessed against the national literacy standards originally designed for native English speakers. This raises interesting questions about the nature of literacy in an increasingly multicultural UK society where plurilingualism is the norm for significant sectors of the population. It also has important implications for curriculum design, testing and classroom delivery as well as for the key measure of the government strategy: the collection of data on achievement. **Szilvia Papp** examines the test that was specifically designed for UK citizenship and settlement, the Life in the UK test, as well as the publicly available materials provided for study towards the test. Using a framework developed by Antony Kunnan and a guide for policy-makers developed by ALTE's Language Assessment for Migration and Integration (LAMI) subgroup in collaboration with the Council of Europe (CoE), she investigates the Life in the UK test for qualities of fairness and validity. She raises the question of how far the language used in the test materials reflects the targeted level of proficiency and language use domain, i.e. the functional competence required for the successful demonstration of citizenship and settlement in the UK.

Section Two of the volume gathers together some valuable insights into the impact of testing and assessment within the more narrowly defined context of language teaching and learning. As our knowledge and understanding of applied linguistics and language pedagogy continues to evolve in light of globalisation, new technologies, changing social needs, and so on, so we can expect to gain fresh insights into approaches to assessing language proficiency, with associated impact on both teachers and learners. **Lynda Taylor**'s opening paper in this section considers the nature of language standards, and the role that the 'setting of standards' and 'standardisation' play in the business of assessment**,** especially for testing agencies. She discusses the reality of

linguistic variation and the implications of this for language teaching, learning and assessment. She identifies the factors (including politics and prejudice) that can shape policy and practice when setting language standards for teaching and assessment, and explores how we might construct a principled and pragmatic approach drawing upon ethical and validation frameworks that have emerged from within the language assessment community in recent years. **John Hawkins and Paula Buttery** report on recent empirical work to investigate the grammatical, lexical and usage competence of English learners at the A2–C2 proficiency levels of the Common European Framework of Reference for Languages (CEFR), which is itself impacting more and more on language education and policy in many parts of the world today. By interrogating Cambridge ESOL's 30-million-word corpus of learners' written responses taken from its English exams (the Cambridge Learner Corpus), it is proving possible to supplement the early, more intuitive profiling of proficiency levels with empirical data on the 'criterial features' that distinguish one level from another. This work, undertaken as part of the English Profile Programme, has potentially significant benefits for both teachers and learners of English, as well as direct application in the construction of well-targeted, level-based assessments. **Wayne Rimmer**'s paper continues this linguistic, corpus-informed theme as he explores the nature of advanced competence, which he suggests is partly characterised by the advanced learner's capacity to play with language and to formulate new patternings or collocations. He defines the ability to select and manipulate a finite language system in order to maximise meaning potential as 'linguistic creativity', speculating that it could offer us a way of recognising and rewarding test takers who are operating at the highest end of the performance spectrum.

Pauline Rea-Dickins, Guoxing Yu and Oksana Afitska broaden our horizons considerably in their paper on the consequences of using an unfamiliar language of instruction (i.e. a second language) within an examination system. They report on the controversial impact that this approach may be having upon the educational progression and outcomes of school-age learners in Sub-Saharan African school systems. They discuss the role of language as a critical factor for effective learning in the African context and consider issues of test fairness and social consequences through the lens of the individual language learner. They go on to reflect upon the potential role of linguistic accommodations for test takers, as well as the effects of classroom language(s) use on learner engagement in subject classrooms and on achievement in formal examinations. **Geraldine Ludbrook** moves us back onto the continent of Europe to explore the certification of teachers' foreign language proficiency in Italy but she maintains the focus on the interface between language competence and subject (or content) teaching. As the provision of Content and Language Integrated Learning (CLIL) moves increasingly into mainstream education, at both primary and secondary level, the call for

certified qualification of CLIL teachers is growing. Her paper reports on a project to design a performance test for Italian CLIL teachers to certify both their L2 competence and their knowledge of CLIL methodology, specifically the teaching of science through English. She discusses how investigation of the target language use through the qualitative analysis of data from CLIL science classroom observation can help to address some of the issues that challenge performance test designers, including construct description and test task development.

Denise Lussier takes us across the Atlantic to Canada and into the relatively unexplored territory of teaching and assessing 'intercultural communicative competence' (ICC). She reports on a study that examined the development of positive cultural representations as an essential component of ICC for better understanding other cultures. Her paper reminds us that language teaching and learning entail so much more than simply the transmission and acquisition of linguistic knowledge and skills. Language education can be an entry-point to cultural knowledge and understanding. It is a discipline which embodies the presence of another culture and contacts with 'the other', and is thus capable of encouraging positive attitudes expressed via behaviours and practices that convey openness to others and other cultures. From North America we move still further round the globe to Japan as **Jamie Dunlea** describes efforts to provide more detailed and useful feedback to learners taking the EIKEN test, a large-scale English proficiency test suite with seven levels. Drawing upon the responses to self-assessment questionnaires completed by passing EIKEN candidates, the construction of the EIKEN Can-do List is designed to create a profile of what Japanese learners at these different levels of ability believe they can accomplish in English in real-life situations. Finally in this section, **Stergiani Kostopoulou** brings us back to Europe offering insights on the potential for using the European Language Portfolio (ELP) to integrate learner self-assessment into the language teaching and learning process. Reporting on a case study from Ireland where immigrant students were receiving English language support in post-primary schools, she sets out to demonstrate the pedagogical and social value of ELP assessment, arguing that integration of the Portfolio into institutionalised language assessment can result in an assessment model that is more educational, democratic, ethical and valid.

Section Three of the volume offers reflections on the impact of tests and testing among the many and various stakeholder constituencies which exist in relation to language assessment, including government policymakers, examination boards, testing agencies, as well as teachers, learners, test researchers and textbook writers. **Micheline Chalhoub-Deville**'s paper takes us back to North America as she addresses the social and educational impact of the US educational reform movement, No Child Left Behind (NCLB) (2001). Within this, she focuses on the standards-referenced assessments (SRAs) that

are intended to measure the progress and attainment of English language learners (ELLs) in terms of *academic* English language proficiency. Her paper explores three primary issues: first, how the L2 construct is operationalised in terms of the SRAs and the extent to which these assessments yield scores that can be interpreted and used to help ELLs; second, the responsibility of test developers and test users with regard to the documentation of impact, and the value of explicit and upfront negotiation of roles, expectations, and activities with regard to impact research; third, the need to expand the traditional conceptualisation of impact research to embrace anticipatory social impact analysis which can inform and guide policy formulation. This perspective is complemented by **James Purpura**'s paper which narrows down the focus to the impact of language assessment on the individual, especially as this relates to individuals engaged in the teaching and learning process. He begins by examining how the research evidence on test impact can be contextualised within a test validity argument for justifying test use. He then explores the research on how large-scale, standardised tests, i.e. those external to the classroom context, impact individuals in the teaching and learning process. Finally, he examines the interface between assessment and second language acquisition, discussing language processing from a cognitive-interactionist perspective and highlighting the role that assessment plays in second language acquisition (SLA) processes. His learning-oriented model of assessment provides a springboard for discussing the potential impact that assessment can have on individual learning in classroom contexts.

Roger Hawkey shifts our attention away from policymakers and learners towards a very different stakeholder constituency in the world of testing and assessment – the constituency of the test materials and textbook writer. He reports on an empirical study commissioned by Cambridge ESOL into the washback of the Certificate of Proficiency in English (CPE) on textbooks designed and used to prepare candidates for the exam. A key aim of this study was to produce evidence in support of validity claims made for the CPE, in particular its *consequential* validity. He describes the research hypotheses, the data collection procedures and the approaches to analysis which enhanced understanding not only of the role of exam-to-textbook washback but also the value of exam washback and impact studies as part of an examination board's research and validation policy.

As we approach the end of Section Three, **Cecilie Carlsen**'s paper takes yet another innovative perspective on the relationship between testing and society, exploring this in terms of the effect *of* society *on* testing, rather than the opposite direction. She illustrates this phenomenon through her discussion of language testing in Norway, more specifically the development and public reception of the national tests of English for Norwegian school children. The final paper in the volume, by **Brian North**, reviews the effect that the CEFR is currently having on discussion of levels and comparison of language learning

outcomes in Europe and beyond. He begins by reminding us of the original purpose and nature of the CEFR and points out the overall effect that it is having on professional networking. He discusses the CEFR's impact on policy development and reports findings from two recent surveys carried out by the Council of Europe's Language Policy Division and from the Language Policy Forum held in Strasbourg in 2007. His paper concludes by considering the practical impact of the CEFR and by assessing its influence on examination reform and on the linking of language assessments in the European context.

With its broad coverage of some key contemporary issues, combining theoretical insight and practical advice, *Language Testing Matters* constitutes a valuable reference work for academics, employers and policy-makers, e.g. universities, education ministries, immigration bodies, throughout Europe and beyond. It will also be a useful resource for postgraduate students of language testing and assessment as well as for practitioners, i.e. teachers, teacher educators, curriculum developers, materials writers, and anyone else who wishes to broaden their understanding of the social and educational impact of language assessment.

Biographical notes on all the volume contributors can be found at the end of the volume, together with a full list of the presentations given at the ALTE Conference in Cambridge in April 2008.

Lynda Taylor and Cyril J Weir
April 2009

Section One
New perspectives on testing for specific purposes

1 When is a bad test better than no test at all?

Rachel L Brooks
Federal Bureau of Investigation
Beth Mackey
Department of Defense[1]

Abstract

Members of the US Government's Interagency Language Roundtable Testing Committee discuss Less Commonly Tested Languages (LCTLs) in the US Federal Government. The article focuses on the issues that arise when testing receptive skills and oral proficiency. The authors discuss how the US Government has adapted its traditional models of test development, validation, and tester training to provide solutions that satisfy both professional standards as well as its own institutional requirements.

Introduction

Gone are the days of the Cold War, when United States Government (USG) language testing organisations focused on a few, well-defined, visible languages, such as German and Russian, with sufficient, advanced resources. Today, intelligence sources come from every corner of the world and in hundreds of languages (NFLC 2005), many of which have never before been tested by the USG. With these new languages come increased challenges to language test development and administration, including difficulty locating appropriate resources, adapting testing instruments to the relevant culture, and standardising procedures across languages with efficient and effective methodology (Collins 2002) in order to accurately report on foreign language abilities.

Language testing is the gatekeeper for all foreign language intelligence and international interests, and the stakes are high for all parties involved (Laipson 2002). Decisions about which personnel are qualified to perform different language-related tasks depend on the validity and reliability of new tests being developed in languages that have never before been needed. Most directly, the quality of such tests affects the careers of the examinees: potential

diplomats, agents, translators, interpreters, etc. USG language testers are charged with identifying these qualified language personnel, and the success of agency missions and the safety of non-language agency personnel depend on their language abilities. In turn, the security of the United States, and in some cases its allies, relies in part on the efficiency and effectiveness of USG agencies' testing practices.

When a USG testing department receives a new language testing request, testing specialists survey available resources and make decisions about how to best fulfil its requirements. Factors such as the timeframe of results, the number of people to be tested, the qualifications and availability of language experts to help develop and administer tests, the funding available for test development, and the applicability of the test to other agencies weigh in on test administration and scoring decisions. As the USG moves from testing in more commonly taught languages to less commonly taught languages, or even almost never taught languages, the challenges posed often mean that test development will be difficult, if not impossible, to undertake. Ultimately, USG testing organisations have no choice; we have to test. The question in the title, 'When is a bad test better than no test at all?' is not fair, because many times there is not a choice whether or not to give a test. The question really should be, 'How do we make the best test possible, given these conditions?'. This paper outlines the steps the USG is taking to produce reading, listening and speaking tests in a wide variety of less commonly taught languages, or perhaps more appropriately, less commonly tested languages (LCTLs).

Background

Many of the initial efforts to develop standardised language assessments in the USG have occurred under the auspices and direction of the Interagency Language Roundtable (ILR) (Herzog 2007). The ILR is an unfunded interagency organisation established for the coordination and sharing of information about language-related activities at the Federal level. It serves as the primary means of communication for departments and agencies of the USG, collaborating on issues of the progress and implementation of techniques and technology for language learning, use, testing, and other related topics. Despite its unfunded status, the ILR has made notable contributions to the language teaching and testing fields from its inception in the 1950s to the present (Chalhoub-Deville and Fulcher 2003, Clark and Clifford 1988, Herzog 2007, Lantolf and Frawley 1985, Lowe Jr 1988). Of particular interest to the language testing field are the Federal Government-wide Language Proficiency Skill Level Descriptions which detail an 11-level scale for foreign language skills of Speaking, Reading, Listening, and Writing (ILR 1985a, 1985b, 1985c, 1985d). Adapted from descriptions originally developed for the United States Department of State in the late 1950s, these ILR Descriptions

have influenced the evaluation of foreign language proficiency both in the United States and internationally (Herzog 2007).

For about 50 years, the USG has used the language testing standards produced by the ILR to develop and conduct foreign language tests to meet the varying demands of its agencies. Initially, the demand was primarily for speaking tests in a fairly consistent set of languages including, but not limited to, Mandarin, French, German, Japanese, Korean, Persian-Farsi, Russian, and Spanish (Clifford and Fischer 1990). Well-educated, highly articulate native speaker testers were identified and methods for training testers were developed (Clark and Clifford 1988). Even in its seminal days, language testing across the Federal Government involved many languages that even today are not frequently taught or tested in academic and commercial contexts. Testing personnel in the USG have considerable experience with these languages and resources for testing in them are typically not terribly difficult to obtain, although recently there has been an increase in the number of tests administered (Tare 2006).

Today, the USG tests in well over a hundred languages for purposes as diverse as measuring the proficiency of diplomats in embassies, interpreters in courts, and soldiers on battlefields (United States 2001). Even though most of the languages sought by the USG today are not taught in the United States educational system, they are essential to operations and lives depend on personnel's ability to reliably use appropriate language skills. Daily decisions are made regarding how to best use the USG's limited resources, given agencies' differing requirements and priorities (NFLC 2005), and language test scores are regularly consulted to make those decisions.

USG agencies conduct tests in reading, listening, speaking, writing, translation, interpretation, listening summary translation, and other skills. Combined across language skills and agencies, the USG administers tens of thousands of language tests each year. Over 12,000 speaking tests alone are administered annually by the Defense Language Institute, the Federal Bureau of Investigation, and the Foreign Service Institute in over 100 languages. Over 500 testers in commonly and less commonly taught languages receive training, attend refresher workshops for re-norming, and undergo quality control checks continually.

In the past, USG test developers and tester trainers had been challenged to find resources in languages such as Hindi, Pashto, Persian-Dari, and Urdu, which still are not commonly taught in the US. Rarely tested by USG agencies in the past, these languages are now tested on a regular basis (Brecht and Walton 1998, United States 2001). Today, a new set of languages, including Baluchi, Sindhi, and Ibo (NFLC nd, United States 2001), impose new demands on USG testing personnel, necessitating adjustments to commonly used testing procedures to conduct language tests under conditions where time is short, resources are scarce, and accurate testing can be literally of

life-and-death importance (The National Language Conference 2004). The development and administration of such new language tests by USG agencies means considering the nature of the language, the language's associated culture, qualifications of language subject matter experts, the language population, and test standardisation. A close examination of these issues, as well as the solutions the Federal Government has come up with to mitigate the problems, will lead to a more complete picture of the changing demands of language testing in the USG (Brecht and Walton 1998, The National Language Conference 2004).

Less commonly tested language issues

Language-specific issues

Anyone who has ever been required to make distinctions between languages has faced the difficult task of deciding what separates a language from related languages, dialects or other variations. Social and political circumstances affect language development, shift, and perception (Gordon 2005). These issues do not just trouble linguistic ethnographers, but also language testers. Test developers have important decisions to make about whether to test variations of a language separately, or as a single language. For example, in the past few years, the USG has moved from testing Serbo-Croatian as one language, to three distinct varieties: Serbian, Croatian, and Bosnian (Gordon 2005). Eastern Punjabi has been separated from Western Punjabi in USG testing. Each decision means redeveloping and validating existing language tests, dividing set resources, and re-evaluating previously established scores and procedures.

Many of the USG traditional testing formats are multi-level tests that include the top end of the ILR scale (Professional, Advanced Professional, and Native, or Levels 3 through 5). Some LCTLs may not be spoken at these higher ILR levels, or if they are, they might be combined with other languages. Some speakers convert to a different language altogether when raising the register, complexity, or sophistication of speech, often a colonial or standard language. Other languages may not simply switch to another language at a certain ILR level, but may switch languages in certain situations or during language tasks, causing the test language to be unable to meet all requirements of a particular level ILR description, as a proficiency test. Speakers of some languages often shift into another language when they move beyond the 'home and hearth' topics. For example, in the Philippines, speakers of Tausug or Chavacano shift either partially or completely into English and Spanish when topics increase in their level of abstraction (Gordon 2005). In some cases, the language does not change completely at the higher levels, but rather adopts a substantial amount of lexicon from another language.

For example Hindi incorporates English, and Cebuano incorporates both Spanish and English (Gordon 2005).

Another issue for consideration is whether every language can fulfil the description of ILR Level 5 without resorting to another language. Some agencies have determined certain language tests to cap off at Level 4, or in some cases, Level 3. Current discussion revolves around Arabic dialect testing, which switches to Modern Standard Arabic in certain contexts (Gordon 2005). Whether or not a single standard can be set for all Arabic dialects is debatable. As the USG increases its testing in the LCTLs, test developers have found the value in first describing the language testing requirement before tackling these issues, as an initial assessment of the purposes of the test results allows the USG to alter traditional formats to test only the pertinent ILR levels.

Determining when language interference is acceptable and when it is not can be difficult, as many languages may not have equivalents for foreign words. Sometimes the adoption of foreign words occurs only in specific subjects, such as technical fields. Language testers must consult with experts in the language to determine a standard for when and where foreign words are acceptable. Receptive skill test developers can avoid some of these pitfalls by omitting any potential test items that may require foreign words. Moreover, they can limit the range of levels that the assessment covers to the lower skill level descriptions, as long as the test fills the need of the relevant agency. Test raters for any open-ended items are trained on acceptable responses.

In productive skill assessment, Oral Proficiency Interview (OPI) testers are trained in how to handle language interference. Examinees are informed before the test starts that they should use foreign words only when they are a part of the language, and to avoid foreign words when target language words are available. If an examinee uses a word from another language or dialect, the tester should ask for explanation in the target language. Problems in communicating are resolved by both the testers and the examinee through circumlocution. Furthermore, periodic retraining is conducted and additional training provided before testing sessions to remind testers of issues of importance, good strategies to employ, and pitfalls to avoid. These sessions often include reminders of the language-specific strategies and language interference issues discussed in previously attended OPI training workshops. Testers are provided with written guides to use during the test, which reinforce the principles and procedures of speaking testing.

Cultural issues

When a USG testing organisation is tasked with developing, administering, and scoring a test in a language not tested before, the test developer not only has to be educated about the nature of the language, but also about the

culture of the land where the language is spoken. Language and culture are inextricably entwined. One sociolinguistic and cultural issue in the target language is taboo topics. Topics that are acceptable for discussion in American culture may be considered offensive or personal in another, and vice versa. This issue needs to be managed very carefully by the tester. Even though test developers may be well-informed and prepared for any test in a new language, issues that present themselves during the course of the tests in LCTLs may be difficult to prepare for beforehand, as tester trainers may have limited knowledge of the language's social and cultural aspects. Experts in the language may know these details implicitly, but may not articulate them to test developers, until they come up in the course of the test.

Testers must also carefully consider the culture-specific appropriateness of particular speaking tasks and role-plays. It is important that the tester trainer be well informed of his or her participants' culture. A mistake like the choice of an inappropriate task or role-play may make the test seem biased or uncomfortable and potentially invalidate the results. Some languages' cultures have gender bias, where there are different expectations for the performance of a male versus a female. In other cases, the power imbalance between males and females may mean that the roles a female can play in a test are limited. As the party who gives the score, the tester has more power than the examinee in speaking tests. When the tester is a female and the examinee is a male, there is an imbalance of power in favour of the woman, which can make the testing experience uncomfortable or unacceptable in some cultures. Female testers from these cultures may find it difficult, if not inappropriate, to challenge men in the process of determining the linguistic ceiling during the course of a test, particularly if the result would be marked linguistic breakdown.

The same principle applies to issues of age and seniority. Younger testers may be hesitant to challenge older examinees, feeling that if they exposed the examinee's linguistic weaknesses, they would show lack of respect for the examinee. Likewise, the more senior examinee may be offended by the challenge from a younger member of the same society, or embarrassed by a weakness displayed during the test. Seniority in employment follows the same pattern. Instances occur when an examinee is a more tenured colleague of the tester. In some cultures, it would not be appropriate to challenge a colleague with more seniority in a way that might lead to embarrassment.

Individual qualification issues

The shift to LCTLs has also necessitated adjustments in the test development training. In the traditional language test development model, practitioners are highly trained, not only in the test language, but also in teaching and testing methodologies. As test development projects have shifted into new

languages, locating qualified, educated speakers of these languages in the United States has proven difficult. Many of the target-language test developers have limited English skills; therefore, training such individuals in the ILR scale and testing models has required USG testing organisations to adapt their traditional models. The USG has found success in pairing highly experienced test development project managers who are adept at dealing with non-native speakers of English with native language consultants, who then work hand-in-hand to develop LCTL tests. This give and take between the language expert and the test developer is time consuming, but has produced successful tests in languages such as Dari and Pashto.

Testers used in LCTLs sometimes, if not often, do not have the ideal profile for language testing projects. They do, however, have the one quality that is irreplaceable, proficiency in the target language. The language testing organisations are challenged to find creative ways to overcome the lack of other necessary qualifications. Some of these tester recruits have no language teaching or testing experience, beyond what they experienced themselves learning a foreign language or undergoing language testing. Tester trainers are challenged to explain how tests are designed and function, and to undo any false perceptions about language testing, such as all native speakers always receive the highest score on the ILR scale. Tester recruits who perform at low levels in English in some or all language skills pose additional complications to trainers. It is sometimes difficult to discern if they internalised the USG standardised language testing system during the training, development, administration, and scoring. If so, to what extent did they grasp the necessary concepts?

Some of the individuals who are recruited have not lived in the country of the target language for decades. The constantly evolving nature of language leads to the possibility that the language as it is spoken in a country today has changed since the native speaker last lived there, and potential testers could be speaking an antiquated form of the language. Moreover, if the tester has been living in the United States for many years, it is possible that lack of practice in the language has caused attrition in the language. Attrition is particularly apparent in the sophisticated or complex speech of ILR Levels 4 and 5 because there are fewer opportunities to use a range of types of speech while living in the United States.

In situations where there is an urgent need and limited resources, USG testing organisations have discovered some options for assistance. Initially, other USG colleagues are consulted on information about the nature of the language, readily available language testing resources or current testing projects underway. Through the work of the ILR Testing Committee, members have developed partnerships across agencies that were not common in years past. A possible solution is to locate readily available materials or a trained, qualified tester at another agency. If none exist, efforts are made to

locate employees within the agency who have the needed language ability on record. Such employees often do not have training in testing, but they usually have a security clearance, a certain amount of availability, and a willingness to help.

If there are no potential testers within the USG, some agencies are free to search for support outside the USG. Language communities in the United States often have community centres and organisations where resources can be found. Additionally, if the languages are taught at the university level, there are professors or other staff who can sometimes lend assistance. USG testing organisations have also tapped professional organisations, which often have a presence on the internet, linking speakers of a particular language who live in various parts of the country. Locating suitable testers is only the first step; the speaker must also be available and qualified to assist with testing.

In some cases, USG test developers and administrators have no choice but to use heritage speakers instead of native speakers. When heritage speakers are used to develop receptive test materials or administer speaking tests, the product may be flawed, threatening the outcome of the test. If the examinee has a higher level of speaking proficiency than the OPI tester, the validity of test results would be affected. Similarly, native speakers have varying levels of language ability, and a native speaker test developer or administrator may have a lower proficiency than the individual taking the test.

When new testers have little background in testing or underdeveloped language skills, individualised training has proven to be more beneficial than training in a workshop setting. Tester trainers can adjust the curriculum to the pace of the trainees, and meet their individual needs. Testers are closely monitored throughout training and test administration for new language-specific or testing theory issues that were not previously addressed, and additional training is provided to address any problems that may arise. If a trainee's native language is deficient in some way, available print and audio media can be used to update language skills. Active testers in all languages are also required to keep current in their language and practice to prevent attrition. Testers are encouraged to practise high-level language skills whenever possible.

In cases where trainees are deficient in English proficiency, an interpreter has been found to be of great use during the tester training. Chances are, though, that an interpreter would be difficult to locate, considering the initial difficulty in finding personnel with the needed language. As an alternative, an interpreter for a language closely related to the target language, perhaps spoken in the same area, can assist in training.

Careful and detailed explanation of what went on during the exam should be documented before assigning a score. Testers review all the tasks and topics posed to the examinee, including responses. Examiners pose detailed questions based on ILR levels, and help the tester to interpret the descriptions.

Only examiners who have an extensive testing and linguistics background should be used in these instances. Only examiners assign scores, though the tester may express an evaluation of what the score should be. Examiners take careful notes on the nature of the language and particular features of the language that were discovered through the course of the test. These notes are used for later reference in conducting tests by the agency and for training new testers in the future. Additionally, the notes can be shared with other agencies. Time and resources permitting, uncertainty in a test score can be resolved through a third party review.

To summarise, the fact that many of the native speakers have not had explicit training in linguistics or language acquisition is further complicated by the fact that all of this information must be relayed in English, which may be an area of weakness for the tester.

Population issues

In testing common languages, testing departments have followed traditional, large scale testing models that rely on piloting, validation and sophisticated item analysis of multiple-choice reading and listening tests. In the LCTLs, USG testing organisations often struggle to find an adequate number of people in the target population who can readily participate in a formal validation. Some of the LCTLs that need testing may come from populations of 10,000 to 100,000 speakers worldwide (Brecht and Walton 1998). The population in the United States to draw from for test development and validation projects is much smaller. In order to collect a large enough sample of speakers, the Defense Language Institute, for example, has had success in including both heritage and native speakers in the validation pool.

Diversifying the language validation pool creates its own set of challenges. Participants may have weak literacy and English comprehension skills; these deficiencies can result in the demands of the test not being met. Since finding large enough populations for thorough item analysis and calibration is difficult, constructed-response tests (CRTs) are being used when testing receptive skills. CRTs are somewhat more direct and flexible than multiple-choice tests, and protocols can be adjusted to accommodate novel examinee responses. The CRT format has been especially helpful in overcoming the difficulties of test developer qualifications and size of the validation population. CRTs are more time consuming to grade than multiple-choice tests, but the flexibility allows for a quicker development cycle. Possible test responses are collected, and using statistical analyses of the most likely and plausible responses, CRT items can be eventually converted to multiple-choice items.

Finding authentic materials in these new languages to be used for reading or listening tests can also be problematic. Media may largely be produced by a diaspora population not representative of the language as it is used

in-country and, depending on the area in question, internet resources may not exist. Some government teams have found limited success in having test developers purpose-write passages for assessment purposes, but care should be taken that the language feels authentic. The USG has also tried to use a variety of diaspora sources, if diaspora sources are the only ones available.

Problems that plague receptive skill testing also affect the testing of speaking. In such testing, some OPI needs may be forecast well in advance, and others at the last minute. While USG agencies have developed capabilities in many languages over the past 15 years, in many other instances, there are no readily available resources, and speakers of that language may be difficult to find. Even when speakers can be found in the US, they may have spent so many years away from their native country that they are not in touch with the language as it is used today. Despite all of these difficulties, the OPI has become the *de facto* emergency language test, due to the fact that an OPI can be administered by trained native speaker via telephone to examinees in remote locations, and a previously prepared form does not have to be developed.

Standardisation issues

USG testing faces the need to create standardised proficiency tests across languages, with emphasis on the Middle Eastern, Central, and Southeastern Asian languages and their dialects. The more experience tester trainers have with testers of various languages and dialects, the better the understanding of how these languages function and interact with each other. Consequently, procedures for speaking testing across languages and agencies must be constantly re-evaluated. In particular, the ILR Skill Level Descriptions need to be applied to each USG test. We have discussed how the nature of these LCTLs is quite different from the languages government agencies are accustomed to testing. As new aspects of these languages emerge, we must interpret the Skill Level Descriptions consistently, to maintain test score reliability and validity. Procedures may evolve to meet USG agencies' changing needs; the language testing ethics and standards cannot be compromised.

Over the past several decades, foreign language test development across the USG has settled into traditional formats. For example, many reading and listening tests are linear, multiple-choice comprehension tests that measure a person's general ability to comprehend spoken or written language regardless of how it was learned, with reference to the ILR Skill Level Descriptions for Reading or Listening. Accordingly, OPIs conducted by USG agencies use the ILR Descriptions for Speaking. Studies of rating consistency have been repeated over the years. Moreover, the regular monthly meetings of the ILR present opportunities for individual agencies to share efforts to tackle pertinent language issues, as well as display advances made in language testing,

training, and translation practices with the foremost language experts. Despite these efforts towards standardisation, the operational demands of the emerging languages do not allow USG testing organisations to meet each standard every time.

Obviously, the first testers of a language in the USG cannot be assessed according to the ILR scale. However, the agencies must present a strong justification based on the credentials presented by the tester recruit, specifically how long and at what age the potential tester lived in areas where the language is spoken. Was that the primary language spoken in that area? To what extent did the individual receive formal education in that language? Agencies investigate how the potential tester used the language. Did the individual use the language in conducting business transactions? Was the language used only at home, or also with friends, neighbours, community members, or colleagues? Did the individual ever give speeches or lectures using the language?

In some cases, the tester recruit has taken other USG tests in that language, but in other skills, such as listening, reading, or writing. Those scores are taken into careful consideration. Scores from tests given outside of the USG are also considered, with attention paid to the standards and ethics used by the organisations administering them. Certifications in teaching, translating, or interpreting the language from the United States or another country, or awards given for work done in the language, and publications in the language become part of a portfolio justifying the suitability of the tester. Nevertheless, this type of evidence is not considered equivalent to the prerequisite testing required of other testers and test developers. As more testing instruments for the emerging language become available, the validity of prior decisions made for language professionals is regularly re-evaluated. Until assessments for the emerging language are considered to meet all the standards previously established for language testing materials, the tester status is considered provisional.

Time also becomes an issue in training. Standards for tester training require testers to go through at least a 10-day intensive workshop followed by many hours of individualised coaching at some agencies, while others require over 300 hours of training before administering official tests. New testers of emerging languages are typically trained for a specific, urgent need, meaning training needs to occur as quickly as possible. When a tester needs to be trained to give a test within hours, standard procedures obviously cannot be followed. Part of every agency's tester training is conducting several practice exams with volunteer examinees off the record. Again, the limitations to the number of speakers of less commonly tested languages restrict the ability of testers to practise before administering a real test. With time limitations, the only alternative is to train the testers as best as possible, and then continue to train them after the first test to prepare them for the next one.

Moving forward

Looking forward, there is much to be done in testing reading, listening and speaking in order to keep up with the USG's demands to test emerging and LCTLs. Continued collaboration not only within the USG and other testing organisations in the United States, but also in other countries will help agencies be increasingly prepared to meet challenges. USG testing organisations can work together with similar foreign organisations to establish speaking testers in new languages by assessing our recruits and informing us about the nature of language. Information dissemination and co-operation will help all language testers to develop valid and reliable testing procedures for new languages quickly and efficiently.

Note

1. The authors would like to acknowledge the contributions of Mika Hoffman, Defense Language Institute, Christina Hoffman, Foreign Service Institute, and Maria Brau, Federal Bureau of Investigation.

References

Brecht, R D and Walton, A R (1998) National needs and capacities: a recommendation for action, in *International Education in the New Global Era: Proceedings of a National Policy Conference on the Higher Education Act, Title VI, and Fulbright-Hays Programs, January 23 – 25 1997,* Los Angeles, CA: University of California, International Studies and Overseas Programs, 93–102.

Chalhoub-Deville, M and Fulcher, G (2003) The oral proficiency interview: A research agenda, *Foreign Language Annals* 36 (4), 498–506.

Clark, J L D and Clifford, R T (1988) The FSI/ILR/ACTFL Proficiency Scales and Testing Techniques: Development, Current Status, and Needed Research, *Studies in Second Language Acquisition* 10 (2), 129–47.

Clifford, R T and Fischer, D C (1990) Foreign Language Needs in the U. S. Government, *Annals of the American Academy of Political and Social Science* 511, 109–121.

Collins J (2002, January 16) [briefing] The National Briefing On Language and National Security hosted by The National Foreign Language Center (NFLC), Washington, DC, retrieved 4 August 2008 from <http://www.nflc.org/policy_and_strategy/language_and_national_security/nflc_briefing_january_2002/full_transcript>

Gordon, R G, Jr (Ed.), (2005) *Ethnologue: Languages of the World*, Fifteenth edition, Dallas, TX: SIL International, online version: <http://www.ethnologue.com/>

Herzog, M (2007) An overview of the history of the ILR language proficiency skill level descriptions and scale, *Interagency Language Roundtable*: Washington, DC, retrieved 4 August 2008 from <http://www.govtilr.org/Skills/index.htm>

Interagency Language Roundtable (ILR) (1985a) Interagency Language Roundtable Skill Level Descriptions for Listening Proficiency, *Interagency Language Roundtable*: Washington, DC, retrieved 4 August 2008 from <http://www.govtilr.org/Skills/ILRscale3.htm>

Interagency Language Roundtable (ILR) (1985b) Interagency Language Roundtable Skill Level Descriptions for Reading Proficiency, *Interagency Language Roundtable*: Washington, DC, retrieved 4 August 2008 from <http://www.govtilr.org/Skills/ILRscale4.htm>

Interagency Language Roundtable (ILR) (1985c) Interagency Language Roundtable Skill Level Descriptions for Speaking Proficiency, *Interagency Language Roundtable*: Washington, DC, retrieved 4 August 2008 from <http://www.govtilr.org/Skills/ILRscale2.htm>

Interagency Language Roundtable (ILR) (1985d) Interagency Language Roundtable Skill Level Descriptions for Writing Proficiency, Interagency Language Roundtable: Washington, DC, retrieved 4 August 2008 from <http://www.govtilr.org/Skills/ILRscale5.htm>

Laipson, E (2002, 16 January) [briefing] The National Briefing On Language and National Security hosted by The National Foreign Language Center (NFLC), Washington, DC, retrieved 4 August 2008 from <http://www.nflc.org/policy_and_strategy/language_and_national_security/nflc_briefing_january_2002/full_transcript>

Lantolf, J P, and Frawley, W (1985) Oral-proficiency testing: A critical analysis, *The Modern Language Journal* 69 (4), 337–345.

Lowe Jr, P (1988) The unassimilated history, in Lowe Jr, P & Stansfield, C W (Eds), *Second language proficiency assessment: Current issues,* Englewood Cliffs, NJ: Prentice Hall Regents, 11–51.

Tare, M (2006) Assessing the foreign language needs of the Department of Homeland Security, *Journal of Homeland Security and Emergency Management* 3 (1), doi: 10.2202/1547–7355.1150.

The National Foreign Language Center (NFLC) (2005, 1 February) A call to action for national foreign language capabilities, retrieved from <http://www.nflc.org/policy_and_strategy>

The National Foreign Language Center (NFLC) (nd) Critical Languages in Southwest Asia, retrieved 5 August 2008 from <http://www.nflc.org/policy_and_strategy/language_and_national_security/critical_languages_in_southwest_asia>

The National Language Conference (2004) *An introduction to America's language needs and resources*, briefing document from the National Language Conference. Conducted by the Center for Advanced Study of Languages and the University of Maryland, College Park, MD.

United States (nd) National Security Education Program (NSEP) Analysis of Federal Language Needs. Congressional Record: 21 March 2001 (Senate) P. S2723 – S2725 <http://www.fas.org/irp/congress/2001_cr/s032201.html>

2 Social, safety and economic impacts of global language testing in aviation

Philip Shawcross
President – International Civil Aviation English Association (ICAEA)

Abstract

Air travel remains statistically the safest form of transport, yet in 1996, a mid-air collision over India resulted in the loss of 312 lives. The investigation showed that inadequate spoken English on the part of the pilots and/or the air traffic controllers had been a contributing factor. These findings spurred ICAO (International Civil Aviation Organization), the United Nations regulatory aviation agency, to set in motion a process which resulted in a new global Language Proficiency Requirement. The ICAO Language Standards represent the framework for the world's first legally binding, global language testing policy affecting a whole industry. Complying with the high-stakes and operationally focused requirements of an aviation English testing process represents a very significant challenge for the language testing industry. This paper will briefly review the impact of aviation language testing from safety, professional, social and economic standpoints.

Background

Language and communication in aviation

A Chinese pilot flying from Beijing to Paris may cross 10 national boundaries and speak to more than two dozen air traffic controllers, each with a different first-language background, speaking different regional varieties of English at varying levels of proficiency. According to international regulations, while pilots may use the language of the country they are flying over, pilots and controllers must be able to communicate in the common language of aviation: English.

Safe flights depend on successful pilot and controller communications. In fact, between 1970 and 1995, accident investigators determined that more

than 1,500 passengers and flight crew lost their lives in accidents in which inadequate English language proficiency on the part of controllers and/or pilots had been a contributing factor. In 1996, a mid-air collision over India resulted in the loss of 312 lives. In this accident, as in others previously, the investigation showed that inadequate spoken English had been a contributing factor.

Most pilot–controller communications employ what is called 'standard ICAO phraseology', i.e. internationally recognised formulaic expressions which are used unfailingly to address routine and foreseeable abnormal situations. Examples of standard phraseology are:

'Cleared for ILS approach Runway 1–3 Right.'

'Start-up approved.'

'Request holding instructions.'

'Leaving Level 3–1–0 for Level 3–5–0.'

'Report leaving Flight Level 3–5–0.'

However, in many non-routine, abnormal or emergency operational situations such as system failures, passenger illness, deviated flights, bad weather conditions, obstacles on the runway, threatening passenger behaviour, running short of fuel, delays, bomb scares etc. standard phraseology is not enough for effective and unambiguous communication. Pilots and controllers must then revert to what is called 'plain' or 'common' language to manage situations. This may include utterances such as:

'The cabin crew have reported three passengers concussed, possibly with broken ribs.'

'We have ordered an ambulance to be standing by at the gate.'

'There seems to be a fuel spillage on Taxiway November.'

'Two passengers are missing and we have had to unload their baggage.'

'We heard a loud thud just after take-off and suspect a bird strike. There was a flock of gulls near the runway threshold.'

Regulating to improve safety in aviation

Accident investigations revealed that it was the use of non-standard phraseology and inadequate proficiency in plain language which were contributing factors in a significant number of aircraft accidents and incidents. As a result, ICAO (International Civil Aviation Organization), the United Nations regulatory aviation agency with 192 member States, based in Montreal, which legislates on every aspect of commercial aviation operations, set in motion a process to study how the level of radio communication could be improved and implemented the necessary measures.

In 2000, ICAO formed the Proficiency Requirements in Common English

Study Group (PRICESG) made up of an international panel of operational and linguistic experts in order to examine the use of English in aviation and make recommendations for regulating it.

In 2003, the ICAO Council approved new Standards and Recommended Practices with respect to Language Proficiency Requirements (LPR) comprising a six-level rating scale and holistic descriptors, defining a minimum Operational Level (Level 4) and establishing the requirement for all pilots and controllers to demonstrate their language proficiency and have their licences endorsed, with the recommendation of periodic re-testing for all those below Level 6. These requirements were scheduled to come into effect in March 2008.

In 2004, the *Manual on the Implementation of ICAO Language Proficiency Requirements* (Document 9835) was published and the first ICAO Aviation Language Symposium was held in Montreal. In 2005, the PRICESG linguistic sub-group met to work on and calibrate recorded speech samples, develop rating rationales and a rating tool entitled the *ICAO Language Proficiency Requirement Speech Sample Training Aid* CD. Over the last four years ICAO has conducted numerous regional seminars to explain the requirements and support the member States in implementing them.

In 2007, a second ICAO Aviation Language Symposium was held and the ICAO Assembly passed a resolution (A36-11) granting the 192 member States the possibility of an additional three-year period to reach compliance – i.e. to make sure that all their international pilots and air traffic controllers handling international flights had reached Level 4 – provided they filed a detailed implementation plan on how this was to be achieved and what contingency measures were being taken in order to ensure safety in the interim period.

ICAO does not possess the internal means or expertise to produce an aviation English test; nor is this part of its mandate. However, in 2008, ICAO commissioned and published *Language testing criteria for global harmonization* in the form of a draft circular for the language and aviation communities. Aviation English training guidelines have been drafted and are scheduled for publication in 2009. Both publications are conceived with a view to offering frameworks to enable the community to regulate itself in these two areas. The concern which has driven all the above measures is to improve safety in what is already a statistically very safe form of transport. These standards represent the world's first language testing policy affecting a whole industry, and naturally present a very complex set of consequences and requirements.

Specific features of the use of language in aviation and the ICAO language requirements

To better understand the specific nature of aviation English testing, it is probably useful to recall the specific features of the language used by pilots

and controllers, and hence the specific nature of testing systems which will be appropriate to assess proficiency in the profession.

Aviation communication is predominantly oral. The aviation English required by pilots and controllers is essentially communicative. Most communication is without any visual contact; even communication between crew members tends to be without use of eye contact or body language with pilots facing forward in a cramped flight deck environment and usually communicating over the intercom with cabin crew. The operational aviation community employs a very specific and varied lexical corpus (weather, mechanics, aerodynamics, security, health, geography, human behaviour, navigation, airport infrastructure, safety etc.), often uses common words in a way which differs from everyday usage ('hold', 'clear', 'advise' etc.) and has a circumscribed range of operationally relevant language functions (orders, requests, offers to act, feasibility etc.) and dialogue management. Aviation radiotelephony communication is typically a blend of formulaic standard phraseology, punctuated by common or natural speech each time a non-routine situation – however trivial – occurs. Moreover, communication is often conducted in a stressful environment where time is a critical factor.

The ICAO Rating Scale covers six language skill areas: pronunciation, structure, vocabulary, fluency, oral comprehension and interactions. Reading and writing are not considered as relevant skills in this context. Even fundamental language competency such as grammar, syntax and vocabulary are assessed more in terms of effective communication in an operational environment rather than in purely linguistic terms. There will typically be more tolerance in certain cases where misunderstanding cannot be generated and greater severity in others where a communication may be equivocal. The ultimate level of language proficiency in aviation (ICAO Expert Level 6) is not native speaker-like English, but a language easily 'intelligible to the international community'. Indeed, by their pronounced accents, use of idiomatic expressions and high rate of delivery, many native speakers may not comply with the criteria of Level 6. Finally, in any ICAO-compliant language test, the various levels of proficiency are defined by the lowest score in all six skills; aggregates are not used.

These features specific to the use of English in aviation have meant that no single existing general purpose test is entirely fit for purpose to assess a person's proficiency in accordance with the ICAO Rating Scale and holistic descriptors, and that it is not possible to establish total equivalence between the levels of any existing scale and those of the ICAO Rating Scale.

However, perhaps even more noteworthy than all the points above is the sensitivity and the safety-critical nature of speech acts in operational aviation.

'We are at take-off.' (KL 4805, Tenerife 1977)

'We are running out of fuel.' (Avianca 052, New York 1990)

These apparently anodyne remarks were part of scenarios which led to two sadly infamous aircraft accidents resulting in the loss of over 600 lives. Within an aviation context, these statements contain a considerable potential for misinterpretation: the first may be interpreted as meaning 'We are ready for take-off' or 'We are taking off'. Fuel reserves are a common issue discussed by pilots and controllers and enter into negotiations for priority during approach and landing; however, in such circumstances pilots need to be able to communicate precisely different degrees of urgency. In both of these cases, the failure or inability of the controller or pilot to paraphrase or challenge the information transmitted led to misunderstandings which proved fatal.

Impact on safety

In aviation operations, where every eventuality is seemingly provided for and yet the unexpected still happens, language is in a very real sense the final safety net. Accidents never have a single cause. To apply James Reason's Swiss cheese model, a whole series of safety barriers has been set up in aviation to prevent an accident occurring and to contain the effects of any failure or human error. Yet every barrier may be pierced. It is when a series of failings in a succession of safety barriers somehow become lined up that the unthinkable happens. Language communication accompanies most of these barriers to make them more effective: pilot to pilot, pilot to controller, pilot to cabin crew. Had the controller working in dense fog in the Canaries in 1977 had greater linguistic confidence and sensitivity, perhaps he would have challenged the Dutch pilot to clarify his statement 'We are at take-off'. He would then have realised that the Dutch plane was not ready for take-off, as he expected and had instructed, but had begun its take-off roll. If the South American pilot in New York in 1990 had been able to make clear by use of paraphrase just how low on fuel they were, and if the American controller had asked for clarification, perhaps history would have been different.

For much of any flight pilots rely almost totally on their ears to acquire what is called 'situational awareness', i.e. knowledge of the environment in which they are flying and into which they will fly: the weather, obstacles and other aircraft. Conversely, air traffic controllers on the ground rely entirely on their ears to know what is happening to and on each flight. Standard ICAO phraseology allows pilots and controllers to manage movements and situations most of the time in the most concise, regulated and unequivocal manner. However, unexpected, non-routine situations often need to be managed using plain language. The importance of effective oral communication is compounded by the growth in the volume of international air travel and the cosmopolitan nature of the staff involved. Emirates Airlines employs some 65 different nationalities in their flight crew. On an international flight,

a pilot will be confronted with controllers speaking English with different accents and degrees of proficiency.

It was in order to enhance safety that ICAO, the world's regulating body for aviation, moved to define standards for the language used by the operational community and enforce the ongoing assessment of language proficiency. To quote from the Holistic descriptors in the appendix to Annex 1 of the ICAO *Manual on the Implementation of ICAO Language Proficiency Requirements* (ICAO Doc. 9835 Appendix A2):

> Proficient speakers shall:
>
> communicate effectively in voice-only (telephone / radiotelephone) and in face-to-face situations; communicate on common, concrete and work-related topics with accuracy and clarity; use appropriate strategies to exchange messages and to recognize and resolve misunderstandings (e.g. to check, confirm or clarify information) in a general or work-related context; handle successfully and with relative ease the linguistic challenges presented by a complication or unexpected turn of events that occurs within the context of a routine work situation or communicative task with which they are otherwise familiar; and use a dialect or accent which is intelligible to the aeronautical community.

The obligation for all pilots flying internationally and air traffic controllers handling international traffic to demonstrate through a testing process approved by their national civil aviation authorities that they have attained at least ICAO Operation Level 4 certainly acts as a powerful force driving up standards of spoken English in the aviation community worldwide. Indeed, in Europe and elsewhere, as regional legislation is brought into line with ICAO requirements, the trend is to aim at higher levels of proficiency. This will mean that pilots and controllers will be better prepared to deal effectively with the potentially hazardous situations with which they may be confronted and that generally speaking there will be greater awareness in the industry of the essential part played by language in the communication-technical-human factors equation. This illustrates how aviation English proficiency testing has probably the highest stakes of all language testing in as much that inadequate or inappropriate testing could result in professionals with poor communicational skills representing a potential hazard for the ever-growing millions of members of the travelling public.

Professional impacts

Since 5 March 2008, with a conditional 3-year period of extension until March 2011, all pilots and air traffic controllers working in an international environment are required to have their professional licences endorsed to

certify that they have successfully passed an approved test demonstrating that they are at least at ICAO Operational Level 4. Without this endorsement, they are unable to work legally in an international environment. Those test takers having attained Level 4 or Level 5 are required to re-sit a test periodically; every three years is recommended at Level 4 and every six years at Level 5. Given that much operational communication is fairly routine and largely based on standard phraseology, pilots and controllers may not often be faced with situations which maintain a wider linguistic ability.

Therefore, success or failure in a proficiency test can determine whether a person retains his or her job or receives promotion: flying internationally, being promoted to captain, flying on larger aircraft, having greater responsibility as a controller, or not. Equally, at the bottom of the professional ladder, it will mean whether a person is hired or accepted for pilot or controller training. This will have immediate and possibly drastic effects on employees' incomes and lifestyles, especially for those already in positions of authority whose initial language training is probably a more distant experience and was maybe less effective.

The testing obligations of the current Language Proficiency Requirements have a direct impact on how airlines and air navigation service providers manage the availability and training of their staff. In a profession where levels of remuneration and social prestige are habitually high, where training is a long and costly process, and where there is currently a shortage of qualified staff, the threat of failing a language proficiency test is particularly acute for both the organisation and the individuals concerned.

Social impacts

The social impacts of the presence of a universal requirement to demonstrate proficiency in order to obtain the endorsement of one's professional licence vary greatly and reflect both varying levels of fluency in the language in different parts of the world and also profound political and cultural differences. In the framework of the present paper, there is only space to refer to a few instances as illustrations of the way in which this unique testing environment affects social behaviour.

Airline captains in particular enjoy considerable status both in society in general and more specifically in their professional environment where their authority over the rest of the crew is considerable. In certain Eastern cultures especially, the distance between captain and first officer (co-pilot) can be extreme, with it being difficult for the junior pilot to question his senior at all. Indeed, much effort is currently being deployed in human factors and Crew Resource Management training to make crews more aware of the potentially dangerous consequences of insufficient interaction between the two pilots. Therefore, a testing system which may suddenly upset this balance

of authority by threatening the licence of the senior pilot, whose grasp of English may well be less robust than that of the junior first officer, can be very destabilising indeed.

Moreover, in certain cultures, failure is not perceived as an option in high-profile professions exposed to the international gaze. Hence the results of certain benchmark tests in one country have been published with a 99.5% pass rate. Similarly, in another country all the questions in the test approved by the civil aviation authorities have been posted on the internet for candidates to become familiar with them thus seriously undermining the validity of the results in both cases. It becomes apparent that the testing systems approved by the national authorities are subjected to intense social and political pressure.

The conditions of compliance with ICAO Language Proficiency Requirements stipulate that States are obliged to file a difference with ICAO if they fail to comply and inform all those States to which their aircraft fly or over whose territory their aircraft fly of their failure to reach Level 4 entirely. Furthermore, language proficiency is henceforth one of the items which are addressed by ICAO safety audits when their officials visit different States. These facts all result in the language proficiency issue becoming a particularly public one. When communication broke down between a Far Eastern flight crew and an American controller late in 2007 it made the headlines on CNN and elsewhere. In the resulting investigation, it appeared that the crew members had been attributed Level 4 and yet in the ensuing interviews it was painfully clear that their communicative ability did not meet the criteria of Level 4.

Airlines are commonly the flag carriers of their respective States and as such have considerable prestige. The thought that their reputation could be tarnished by the negative publicity of having some of their staff declared non-compliant with the Language Proficiency Requirements can understandably generate great anxiety and threaten national pride. This in turn may lead to pressure to design or adopt a testing system in which failure is marginalised or results tampered with. Consequently, in a testing process in which the stakes are so high from many points of view rigorous security is a paramount, but not inviolable, issue. In other countries, labour laws or strong trade unions may prevent dismissal or reclassification and characterise as discrimination the testing and sanctioning of professionals hired at a time when language requirements were not in force thus appealing to 'grandfather laws'. This in turn may result in either testing being disregarded or the unrealistic training of more senior staff.

Finally, on a more personal level, failure or the fear of failure in a language proficiency test, which determines them practising their profession and maintaining their livelihood, may affect the self-esteem of otherwise highly qualified and respected professionals in their thirties, forties or fifties. This

adds to the anxiety experienced in professions already regulated by frequent medical check-ups and professional checks.

Economic impacts

The implementation of the ICAO Language Proficiency Requirements comes at a cost to the aviation industry as a whole, both in direct and indirect expenses, which has been roughly estimated to be at least in hundreds of millions of dollars. In addition to the expense of procuring and administering generally sophisticated and secure purpose-built tests for hundreds of thousands of pilots and controllers on a recurrent basis, there is the considerable cost of taking highly paid professionals working on rosters or shifts off the job in order to sit the test and then possibly follow extensive remedial training if they fail it. Any failure to pass the test will in turn result in the employee being withdrawn from international operations, or being subject to conditional contingency measures until 2011 (e.g. working necessarily with a colleague whose licence has been endorsed), in a work market where there is already a severe shortage of qualified professionals and financial pressures due to fluctuating fuel costs and a more vulnerable global economy. The Language Proficiency Requirements are indeed challenging for Human Resource departments.

At a higher level, in a global economy with an intensely competitive market where travellers select the airline they use online, any negative publicity about the language proficiency of a given airline's staff or a country's controllers has the potential to have a very detrimental effect upon their image in the eyes of the travelling public and hence have a direct impact on ticket sales. A time is dawning when language proficiency is entering people's awareness as one of the parameters to be taken into account in air travel along with safety records, fares, punctuality, quality of service, ease of connections, leg room, baggage handling and on-board meals.

Conclusions

Has ICAO opened a Pandora's box in creating requirements for global testing in aviation? It is time to recall a few facts which may help place things in perspective. Like the railways in the 19th century on a national level, aviation has accelerated a standardised awareness of time and space at a planetary level and been one of the main motors of technological and economic change. Whether for better or for worse it is an integral part of our world system. In the process, aviation has had some far-reaching effects on society.

Statistically, aviation remains the safest form of travel. More concerted efforts and funds are devoted to improving safety and security in aviation than in any other field of human activity.

Not only has aeronautical technology driven research which has had important beneficial effects in other areas, but the aviation industry has pioneered an awareness of and research into human factors and team resource management which has directly benefited practice in operating theatres, high-speed trains and nuclear power stations.

In the final analysis, the objective of the Language Proficiency Requirements is safety. It is beyond doubt that in a multi-cultural, yet highly regulated, society with an exponential growth in air travel, enhanced and more reliable communication is a vital component of a safer world.

To conclude by returning to a standpoint of applied linguistics and language testing, the case of global language testing in aviation can be seen as having singular significance. Not only may it suggest a model for the way in which our society feels it needs to ensure linguistic competence beyond an academic context in specific areas of activity where the accurate and reliable use of language is critical, using purposefully designed assessment tools, but it is also indicative of our more developed awareness of the essential role effective oral communication plays at the heart of an increasingly complex and technological world.

References

Aviation English

Albritton, A (2007) ICAO Language Proficiency in Ab-initio flight training, *Second ICAO Aviation Language Symposium*, Montreal: www.icao.int

Alderson, C (2008) Report on a survey of aviation English tests. www.icaea.pansa.pl

Alderson, C and Horak, T (2008) Survey of national civil aviation authorities' plans for implementation of language proficiency requirements. www.icaea.pansa.pl

Cushing, S (1991) Social/cognitive mismatch as a source of fatal language errors: implications for standardization, *Fourth ICAEA Forum on Aviation English Standards*, Paris: www.icaea.pansa.pl

Cushing, S (1995) Pilot-Traffic Control Communications: It's not only what you say, but how you say it, *Flight Safety Digest*, July 1995.

Day, B (2004a) Heightened awareness of communication pitfalls can benefit safety, *ICAO Journal* 59 (1).

Day, B (2004b) ICAO Standards and Recommended Practices – an overview, *First ICAO Aviation Language Symposium*, Montreal: www.icao.int

Fox, M (2007) Language Proficiency: Implementing the Requirements, (ppt) *Second ICAO Aviation Language Symposium*, Montreal: www.icao.int

Gault, I (2007) Aviation English, *Eighth ICAEA Forum on Aviation English Training: choices & solutions*, Cambridge: www.icaea.pansa.pl

Green, E (1991) The enforcement of RTF phraseology and aspects of callsign confusion, *Fourth ICAEA Forum on Aviation English Standards*, Paris: www.icaea.pansa.pl

ICAO (2007) Implementation Checklist, *Second ICAO Aviation Language Symposium*, Montreal: www.icao.int
Mathews, E (2004a) New provisions for English language proficiency are expected to improve aviation safety, *ICAO Journal* 59 (1).
Mathews, E (2004b) The role of language in aviation communications, *First ICAO Aviation Language Symposium*, Montreal: www.icao.int
Mathews, E (2004c) ICAO language proficiency requirements, *First ICAO Aviation Language Symposium*, Montreal: www.icao.int
Mathews, E (2007) The value of content-based language training for the aviation industry, *Second ICAO Aviation Language Symposium*, Montreal: www.icao.int
McGrath, M (2007) Sharing resources for English language improvement in international aviation, *Second ICAO Aviation Language Symposium*, Montreal: www.icao.int
Mell, J (2004a) Language training and testing in aviation need to focus on job-specific competencies, *ICAO Journal* 59 (1).
Mell, J (2004b) Specific purpose language teaching and aviation language competencies, *First ICAO Aviation Language Symposium*, Montreal: www.icao.int
Mitsutomi, M (2004) Some fundamental principles of language teaching and learning, *First ICAO Aviation Language Symposium*, Montreal: www.icao.int
Mitsutomi, M (2005) Language acquisition, *Seventh ICAEA Forum on Teaching and Learning Aviation English*, Besancon: www.icaea.pansa.pl
Mitsutomi, M and O'Brian, K (2004) Fundamental aviation language issues addressed by new proficiency requirements, *ICAO Journal* 59 (1).
Shawcross, P (2004a) Proficiency requirements underscore importance of teaching and testing, *ICAO Journal* 59 (1).
Shawcross, P (2004b) Technology in language teaching, *First ICAO Aviation Language Symposium*, Montreal: www.icao.int
Shawcross, P (2007) What do we mean by the "washback" effect of testing? *Second ICAO Aviation Language Symposium*, Montreal: www.icao.int

Aviation background

Beaty, D (1995) *The Naked Pilot: the human factor in aircraft accidents,* Shrewsbury: Airlife Publishing.
Cushing, S (1994) *Fatal Words,* Chicago: University of Chicago Press.
Duke, G (1998) *Air Traffic Control*, London: Ian Allan.
Godwin, P (2004) *The Air Pilot's Manual*, volumes 1–7, Cranfield: Air Pilot Publishing.
Henley, I (2003) *Aviation Education and Training,* Aldershot: Ashgate Publishing.
ICAO (2000) *Human Factors Guidelines for Air Traffic Management (ATM) Systems, ICAO Document 9758-AN/966,* Montreal: ICAO.
Isaac, A and Ruitenberg, B (1999) *Air Traffic Control: Human Performance Factors*, Aldershot: Ashgate Publishing.
Kirwan, B, Rodgers, M and Schafer, D (Eds) (2005) *Human Factors Impacts in Air Traffic Management,* Aldershot: Ashgate Publishing.
Marriott, L (1990) *From the Flight Deck 3: BAe 146 in Europe*, London: Ian Allan.
Nevile, M (2004) *Beyond the Black Box,* Aldershot: Ashgate Publishing.

Reason, J (1990) *Human Error,* Cambridge: Cambridge University Press.
Stewart, S (1984) *From the Flight Deck 1: Heathrow Chicago*, London: Ian Allan.
Wild, T (1996) *Transport Category Aircraft Systems*, Englewood CO: Jeppesen.

Official bodies

Flight Safety Foundation (FSF) www.flightsafety.org
International Air Transport Association (IATA) www.iata.org
International Civil Aviation Organisation (ICAO) www.icao.int
International Federation of Airline Pilots Associations (IFALPA) www.ifalpa.org
International Federation of Air Traffic Controllers Associations (IFATCA) www.ifatca.org

ICAO publications

ICAO (2001, 2003) Doc 4444 *Air Traffic Management*, Montreal.
ICAO (2004, 2009) Doc 9835 *Manual on the Implementation of ICAO Language Proficiency Requirements*, Montreal.
ICAO (2006a) *Language Proficiency Requirements Rated Speech Sample Training Aid CD,* (Order No AUD001 – ISBN 92-9194-655-9), Montreal.
ICAO (2006b) Doc 9432 *Manual of Radiotelephony*, Montreal.
ICAO (2008) Circular 318 – AN/180: *Language testing criteria for global harmonization.*

3 Going from language proficiency to linguistic evidence in court cases[1]

Margaret van Naerssen

Immaculata University

Abstract

With the increase in court appearances of non-native speakers (NNSs) of the primary language of a legal system, and with the growing application of linguistics in legal cases, language assessment procedures and subsequent findings are increasingly being introduced, hence more visible. As such they are increasingly targets of challenges. Thus, those concerned with use of language assessment need to be better prepared to make informed decisions in the use of particular protocols and to evaluate the choices, findings, and use of results by others. Use of evidence argumentation (Mislevy 2003 as presented by McNamara and Roever 2006) is one tool for evaluating and strengthening the links between language proficiency assessment and linguistic evidence.

Introduction

In court cases involving non-native speakers (NNSs) of the primary language of a legal system, it is not enough for language assessment experts to report levels or scores on a language assessment protocol. Language assessment experts, linguists, attorneys and judges need to see beyond the numbers, scales and labels. Language assessment experts understand this basic principle; however, those working with the legal system as well as some linguists may not understand. This is becoming increasingly a concern as in recent years assessment experts are more and more being asked to do language assessments and report their findings in court contexts and other legal settings. As a result, their procedures and findings are increasingly targets of challenges. However, these experts are frequently not familiar with the special conditions involved in language assessment of an individual subject in legal settings.

The purpose of this paper is then to assist these experts in becoming better prepared to make informed decisions in the use of particular protocols and to evaluate the choices, findings and use of results by counter-experts. Of

particular concern is evaluating and strengthening the links between language proficiency assessment and linguistic evidence. Ultimately, assessment experts also need to make their reports and testimony understood by other linguists and by attorneys, judges and members of juries.

In this paper the issues are first introduced through a case study.[2] Key concepts are then reviewed: language proficiency, language samples/evidence, and linguistic evidence. The concepts are reviewed especially in the context of NNSs. Several relevant questions emerge:

- Do all language proficiency assessments actually constitute appropriate sources of linguistic evidence?
- Is such linguistic evidence linked appropriately to the legal issues and relevant communication tasks?
- To what degree might an assessment protocol be vulnerable to manipulation by the examinee, thus, possibly not a source of evidence of true proficiency?

These issues have been addressed before in some individual cases. However, with the increase in NNSs of the primary language of a legal system, the issues need to be more clearly defined and appropriate approaches need to be identified. Language assessment experts and others involved in forensic linguistics need to share their ideas and research. For this paper selected theory and research have been identified relevant to language assessment validity and from work on assessment/evidence. Relevant issues are then applied in one aspect of an actual (but anonymised) case.

Forensic linguistics is a growing area of applied linguistics in which theory, methods, and research from various specialisations of linguistics are applied in legal settings, including language development and assessment. Forensic linguists do analyses of actual evidence in ongoing ('live') cases and do research on language data from completed cases and from a range of law enforcement settings. The focus in this paper is on the issues faced in 'live' cases.

Illustrative case: perjury and fraud

Below is a case to illustrate how language proficiency assessment and conversation analysis can be combined to address the legal issue of perjury in an insurance fraud case. It is not enough to simply provide results of external language assessment. The findings need to be related to the language evidence, to the legal issues, and to the specific context of the alleged criminal activities. Some details are provided to alert language assessment and other experts in linguistics to the precautions that can be taken to reduce the vulnerability of the assessment findings to attack in the court. Space does not permit a full case report. Nevertheless, these basic issues are also explored more later in the paper.

Case summary: perjury and fraud

The defendant, Mr K, went to his insurance company to file a claim to cover the costs of repairing his roof due to damage during a snow storm. Mr K is a non-native English speaker with relatively limited English speaking skills. He took no interpreter and no legal counsel. The visit turned into a formal interview with the insurance agent and with a legal reporter transcribing the interview, of about 1½ hours. Apparently Mr K did not realise the interview would become the basis of perjury and fraud charges. No audio recording was made of the interview. The insurance company reported the situation to the anti-fraud unit of the local police department. The police followed up with two surprise home-visits.

Legal question and issues

> *Is it likely that the immigrant, Mr K, lied in his insurance claim and interview in order to collect money for roof damage repairs resulting from a snow storm on February 5? If it is likely that he did lie – and if the lie was related to information about the claim – then he would be found guilty of both perjury and fraud.*

The defence counsel focused on critical questions on specific pages of an interview transcript. As the defendant was a non-native speaker of English, the defence counsel had to be well prepared for possible attacks by the prosecution. Thus, he expanded his concern: Might the defendant have pretended a lower proficiency level of comprehension and speaking? Was he possibly using avoidance or malingering (pretending a false condition) in addition to lying?

Perjury is generally known as lying when one has sworn to tell the truth in court. In the US context legally there are four conditions that need to be met for someone to commit perjury:

1. Did the person understand the question(s)?
2. Did the person intend to deceive?
3. Did the person actually try to deceive?
4. Was the deception related to another charge in the case?

Language evidence

The initial language evidence consisted of: 1) two police reports of home-visits by the police fraud unit, and 2) a 79-page transcript (about 1 ½ hours) of an interview between the defendant and his insurance agent. The interview was conducted as if it were a cross-examination in court or police interview with questions moving back and forth across various times. This is sometimes

done to try to catch the interviewee on inconsistencies, or more positively to check for accuracy of facts. The police reports were handwritten summaries of the events in the home-visits. The typed insurance interview transcript was turned over to the police as evidence of fraud. While it was not a legal deposition, the court allowed it to be treated as such, thus, alleged lies in the interview would be treated as evidence of perjury.

As a linguistics expert consultant I was asked initially to look at a few pages of one police report to determine whether or not language proficiency might be an issue. While I agreed there was the possibility, I felt that additional direct language evidence of his language proficiency would be needed. Remember that there was no audio recording of the interview. As the existing language evidence was based on oral interactions with the police and with an insurance agent, an oral proficiency interview seemed to be the most appropriate (a modified version of the ACTFL (American Council on the Teaching of Foreign Languages) Oral Proficiency Interview).[3] I did not look at any of the other language evidence before conducting a language assessment interview as I did not want to be influenced by further language evidence: I wanted the interview to be an independent language assessment. Care was also taken not to discuss any aspect of the case during the interview. Also, unlike most language assessment situations, I had to also assume the possibility that the speaker might be faking a lower than truthful language proficiency.

My assessment was that the defendant was performing at Novice High/ Intermediate Low level. This assessment and the audiotape of the interview became a third set of language evidence. A second oral proficiency interview by another examiner might have further strengthened the assessment procedures, but for a couple of practical reasons this was not done.

Triangulating data from language evidence

The defence counsel only wanted me to take the time (reducing his costs) to examine questions on certain pages. I disagreed with the proposed scope as I felt the short interactions just on certain pages would not provide enough language for determining consistency or inconsistency with the language assessment data. I felt the defendant's interactions in other parts of the insurance interview also needed to be examined to determine whether the defendant's language performance was consistent – or whether on the specific pages there was a sudden change in performance. I had to triangulate the data from all three sets of language evidence: insurance interview transcript, language assessment interview, and police reports.[4]

Conversation analysis was used as the primary tool in the analysis of the transcript to determine whether the persons involved in the conversation actually understood each other. The impact of the language proficiency of the defendant also needed to be considered in the Conversation Analysis.

While time (and digital technology available at that time) did not permit an analysis of 100% of the transcript, a sampling of about 42% of the transcript was done with some attention paid to selecting certain types of interactions, including instances of apparent communication breakdowns.

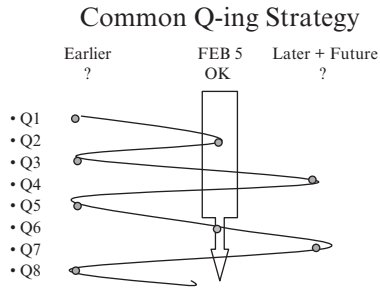
Linguistic evidence – findings

Space does not permit presentation of the details of the analyses. The focus here is on the general issue of how language proficiency is relevant to language evidence. First, across language samples the Novice High/Intermediate Low levels on the formal oral interview were consistent with interactions documented in the interview transcript and with information from police reports from home-visits. Second, within the interview transcript the communication skills, as represented on key pages of the transcript, were generally consistent with skills in the rest of the transcript. Specifically, in the analysis and examination of his second language (L2) skills, it was shown that it was highly likely that he did not understand the formal language used in much of the questioning. He had difficulty with complex questions, including the critical questions asked by the attorney on the 'critical pages' of the transcript. There was also general consistency in terms of developmental patterns.

Third, there was evidence that Mr K understood questions when the time was clearly 5 February or when another time was *specified.* However, for unspecified times in the past or after 5 February or in the future he appeared to answer as though the time was 5 February or he signalled confusion. Suddenly I realised he might be using 5 February as an anchor, and the alleged 'lying' responses then made sense. He very well could have been answering truthfully for *what was 5 February in his mind.* I then re-examined the data in light of this possibility.

Figure 1 illustrates the zig-zag questioning pattern used by the insurance agent in this case as well as that commonly used in some legal questioning

Figure 1 Common Questioning Strategies



(also presented in van Naerssen 2005). It does not represent the actual data, but is similar to a more detailed graphic used in the expert report. The question-marked columns 'Earlier' and 'Later + Future' indicate where there was evidence of communication breakdowns between Mr K and the insurance interviewer. In these situations specific times or dates were also missing. However, when the time was clearly specified, and especially when it was clearly 5 February (date of the snowstorm causing the damage), Mr K appeared to understand and answer effectively.

Linking linguistic evidence to the legal question

Recall that the initial ground for perjury is whether the person understood the questions. While no-one can get inside of another person's head for absolute verification of what the person was thinking, linguistic analysis does provide one window into a person's thinking. It can help the potential fact finder (in this case the judge) to better understand patterns in the language evidence. In this case evidence of a low level of second language proficiency was consistent with other language evidence (in the insurance company interview transcript and at a general level with the police reports). Thus, I was then able to show that it was *highly likely* that he did *not* understand the key questions accurately. Therefore, the other grounds for perjury fell away. Thus, fraud could not be proven.

Key concepts

In the case just discussed, key concepts were mentioned: *language proficiency*, *language evidence*, and *linguistic evidence*. While readers of this paper have a working definition or understanding of language proficiency, the difference between *language evidence* and *linguistic evidence* may seem less clear. These terms are also sometimes used interchangeably even by forensic linguists. However, rigour is required to work logically with evidence argumentation – as required by the law and in language assessment theory. To try to assure a common understanding among legal practitioners, these concepts are reviewed below (from van Naerssen 2007a in Ramirez 2007).

Language proficiency

While there are other sources for definitions on *language proficiency*, for experts working with legal practitioners in the US, this 'government' definition has some legal credibility. An expert needs to find the most appropriate source of a definition for a specific context and a definition that the expert can support. In this definition 'student' can refer more generally to 'learner' or non-native speaker.

> *Language proficiency* 'refers to the degree to which the student exhibits control over the use of language, including the measurement of expressive and receptive language skills in the areas of phonology, syntax, vocabulary, and semantics and including the areas of pragmatics or language use within various domains or social circumstances. Proficiency in a language is judged independently and does not imply a lack of proficiency in another language' (Office of Civil Rights, US Department of Education).

As the focus of this paper is on assessment and evidence, space does not permit developing related acquisition/developmental issues. However, these two critical concepts are important to keep in mind. First, language proficiency *cannot* be measured in inches or centimetres like a line (Underhill 1987). Second, humans are complex, therefore, language is complex. This makes the work of experts dealing with non-native speaker cases, linking language assessment to evidence, especially challenging.

While batteries of tests for different skills/task types are typically used in some second/foreign language assessment, in the courts it appears that generally a single test is being used, though sometimes inappropriately. Efforts to research and evaluate language proficiency and performance in legal contexts should also involve appropriate multi-forms of methodology and assessment.

Language sample and language evidence

In an assessment context, a language sample refers to any oral or written language that becomes the subject of analysis as an insight into the language proficiency of an individual. Language samples can range from natural communication under unmonitored, spontaneous conditions to language produced under tightly controlled conditions and a very limited length, usually with focus on a specific feature. In a forensics context, language samples are the pieces of *language evidence* typically including products of alleged criminal activities, products of investigation, and products of legal processing of suspects/witnesses. A covert recording of communications might produce a sample of naturalistic data as long as the target is *not* suspicious that a recording is being made. Such language evidence can sometimes also be used to gain insights on language proficiency. An oral proficiency interview or a law enforcement interview might be placed somewhere between naturalistic data and highly form-focused assessment, closer on a continuum to naturalistic. It is understood that one variable is the power dominant gatekeeper function of the interviewer. Highly form-focused assessment is unlikely to be considered a valid assessment for a person's communicative abilities in law enforcement settings, although in some circumstances it might be used as a supplementary source of data if certain structures are critical in the case.[5]

Linguistic evidence

Linguistic evidence has been referred to by numerous linguists working in forensic contexts. (Some references are given at the end of this section.) The concept of linguistic evidence is presented to show its relation to and contrast with language evidence, and then to contexts involving NNSs.

The findings of analyses or language assessment, grounded in principled linguistic theory and practice, become *linguistic evidence*. However, other useful information can be found in language samples resulting from content, behavioural and other types of analyses. For example, an eye witness account might contain descriptions of sensory perceptions – what was heard, what was smelled. Although the information is conveyed *through* language, *content* analysis is the tool. This may then provide specific clues to experts in the analysis of the truthfulness of eye witness accounts. Thus, just because an analysis is based on a *language* sample *does not* mean it is *linguistic evidence*. This distinction is not always carefully maintained even by linguists sometimes. Additionally, anecdotal evidence by a non-linguist reflecting on impressions of a person's language abilities (recent or from a time distant in the past) is *not linguistic evidence* (van Naerssen 2007a) .

There is a growing literature on legal cases illustrating specific legal and linguistic issues and analyses from a wide range of specialisations in applied linguistics.

In some cases the studies also underscore the lack of adequate language evidence, but where legal judgments are still made about language meaning and use. (A sampling of collections of studies include Coulthard and Johnson 2007, Gibbons 2003, and Cotterill 2002.) Roger Shuy, in his numerous articles and books, conceptualises linguistic evidence and demonstrates methodologies for the analysis of language evidence in a wide range of contexts. Those concerned with second language and second dialect speakers in the legal system are urged to read Diana Eades' 2003 review of critical issues across a wide range of contexts.

In many studies the assumption seems to be that as linguists we know what linguistic evidence is and that in court we can present analyses and evaluate counter-evidence. However, the distinction frequently is not made between language evidence (as the relevant and legal language sample) and linguistic evidence (product of linguistic analysis). Debate still occurs about what is/is not linguistic evidence, even among forensic linguists, primarily online and at conferences. Shuy has repeatedly stressed the importance of grounding one's analyses in principles from linguistics and using the appropriate tools from the relevant specialisations in applied linguistics. Only then can one claim to be doing forensic linguistics. He cautions against the misapplication of 'linguistics' to 'document analysis, handwriting analysis, type-token analysis, or statement analysis, which come closer to content analysis of words or parts

of words'. He would also exclude certain types of stylistics analyses (Shuy 2006:4–9).

Selected theoretical and practical concerns in the court

This section focuses on four areas: (1) myths and misunderstandings related to language assessment; (2) language assessment and its relationship to linguistics; (3) traps related to truthful language performance in evidence; and (4) a few traps to avoid as an expert. In discussing these myths, misunderstandings and traps, the three questions mentioned in the introduction are explored.

Myths masking validity and other misunderstandings

Language assessment experts interacting with law enforcement, attorneys, judges, and even linguists (without a language assessment background) need to be prepared to address a few myths about language assessment that an expert might encounter especially in the court.

Myth 1: Language assessment is a fixed, time-honoured way of determining proficiency, and this would apply equally in legal cases.

Myth 2: This is a test of language so why should there be an issue?

Myth 3: Language assessment might be seen as an independent professional/academic discipline – so just import the language scores.

Myth 4: Only widely recognised tests are the best to use in legal cases.

There is a major underlying flaw in such myths. They overlook the need to clearly relate language proficiency and other related language evidence to legal issues and contexts. In testing theory this is clearly a validity issue. Continuing with the validity issue, widely recognised tests (designed for large and somewhat diverse populations) frequently do not assess the communicative skills involved in a specific law enforcement or alleged criminal context or involved in possibly producing or understanding other language evidence. These situations also usually involve one person (a single subject) as the focus of assessment.

Courts in the US (at federal and state levels) have rules of evidence that require principled or scientifically based methods and methods or instruments accepted by the professional field. For judges (the gatekeepers on what evidence can be presented in court), reliability rates are 'easy' indicators of 'scientific' status. This may create a dilemma for the language assessment expert who may feel the pressure to privilege reliability over validity. Appropriately the expert may feel that no matter how reliable a test is, if it

is not valid, it is worthless. Thus, the expert needs to be prepared to justify assessment instruments/protocols used, and especially when validity becomes very important.

This dilemma is then a further argument for multiple forms of language assessment. One solution may be first to administer one test that meets the 'widely accepted' criterion, is supported by research, and is somewhat close to the skills involved in the case even if it does not have strong validity.[6] Then balance this off with carefully designed relevant language assessment tasks, grounded in research and that have more validity. In addition, it is wise to gather as much information on the communication skills of the defendant or suspect in the case from other outside sources such as driver's licence tests, educational history and school records, vocational aptitude tests, etc. (but without personal contact with the defendant/suspect). This all needs to be framed in simple terms so that when presenting the findings the expert is also educating the various legal practitioners and the jury.

In addition to the myths mentioned above, there are misunderstandings on the nature of the language of NNSs by legal practitioners, by jurors, and even by some linguists without a background in second language acquisition/development. Research in this area is available in many professional publications. Some generalisations and implications for forensic contexts can be found in van Naerssen (2007a).

Other misunderstandings and traps

Three other areas of misunderstandings involve (a) the basis for linguistics in language assessment, (b) truthfulness about language proficiency, and (3) special court constraints. Awareness of these issues may reduce the vulnerability of an expert when testifying in court or when submitting a report.

Language assessment and linguistics

The views reflected in the previously mentioned myths also ignore the underlying linguistic connection that principled language assessment has with linguistics. Principled language assessment builds on various areas of linguistic research as well as evaluation and statistics. Such language assessment is a sub-discipline of applied linguistics. Being able to present one's field and justify its relevance to a specific case is very important for the credibility of an expert. In the US it cannot be assumed that a judge will automatically accept testimony from a particular professional discipline just because, for example, it has departmental status at the expert's university.

Truthful level of language proficiency

In addition to the validity issues already discussed above, there is also the very important issue of truthfulness of a person's language proficiency in

performance on tests and in other communications that become language evidence. Frequently claims about a person's language proficiency might be made in preliminary motions in order to disallow certain evidence based on a low level of second language proficiency, e.g., 'I didn't understand what the police officer said'. These claims should involve some language assessment support which needs to be carefully done by the assessment expert. Any assessments presented by one side also need to be carefully evaluated when an expert is testifying on the opposing side. Second/foreign language professionals (in language assessment as well as in language teaching) are more frequently concerned about students who might cheat to produce a higher score and are less accustomed to thinking in terms of an examinee faking a lower proficiency for perceived legal advantages. Thus, they might not be tuned into this issue.

In presenting language assessment data in court an attorney the expert is working with may choose to introduce, early on, the issue of potential faking of a lower proficiency just to get it out of the way. Or it may not be mentioned until the cross-examination. The expert should be prepared to show how this potential factor was taken into consideration. The expert should also be alert to hypothetical questions in this regard such as 'Let's assume the defendant was faking/pretending not to understand very well'. 'Well, just for the sake of argument, if the defendant was faking a lower proficiency, wouldn't you say. . .?'. Resist answering such questions. Such questions will quickly draw the expert into a trap. The response should be more along these lines. 'No, I cannot and I did not make such an assumption.' 'Making such an assumption would not be objective.' 'I considered equally both the possibility of truthful and untruthful performance.'

Court constraints on experts

Finally, experts in court need to clearly understand their roles as defined in the law. This varies by country and by levels of a court system. A number of concerns could be discussed, but two basic ones are critical to keep in mind. First, in the US an expert may not testify on content that is *common knowledge*. Defining *common knowledge* is challenging since we all speak a language, thus, we all may feel we have an opinion on language. The expert needs to show the court (the gatekeeper, the judge) what *special* expertise the expert brings that can assist the fact finders. The expertise also must be well-grounded in principled analysis and research. Second, while the expert is allowed to give a professional opinion based on an analysis of the evidence, the expert needs to be careful about *not* drawing conclusions. The expert only *assists* the fact finders (judge and jury) in understanding the language evidence. The fact finders draw the conclusions, make decisions regarding guilt or innocence. These and other court constraints and alerts are discussed in more detail in such sources as Shuy (2006), Coulthard and Johnson (2007), and van Naerssen in Gueldry (in press 2009).

Involvement by the language assessment community

Where is the language assessment community in terms of providing research and other support for language proficiency in legal cases? The language assessment field is, understandably, primarily concerned with assessment for relatively large populations for institutional placement and admissions and other gatekeeping purposes to show growth or predict potential performance beyond the test. However, the somewhat unique issues of language assessment of NNSs in legal systems and related language evidence have not yet received much attention. Nevertheless, the language testing community is a source of valuable expertise for future work in this area. In Australia, the UK and the US (and perhaps other countries) experts in assessment have testified in individual legal cases. Some colleagues in the International Language Testing Association (ILTA) have expressed interest in such forensic applications. While their professional priorities have generally been elsewhere, they have been willing to assist in exploring the issues of linking language proficiency and language evidence to legal issues.[7]

The international testing community has also been growing in its commitment to examining the social impact of language testing. A few examples are the ILTA Code of Ethics and Code of Practice, the focus of the 2008 ALTE 3rd International Conference on language assessment and social impact, and as highlighted in McNamara and Roever's 2006 book and their pre-convention workshop at the 2007 Language Testing Research Colloquium. Their focus on areas of socio-cultural aspects of testing, including validity and testing and identity and the work of others as well in qualitative assessment may also provide useful perspectives in future work in assessment in forensic contexts. In the next section some specific applications to forensic contexts from current language assessment theory are discussed.

Arguing evidence

Like the law, sound evidence-based argument is required in the development of language assessment instruments. Like the law, evidence in language assessment should be relevant to inferences or claims made about the examinee (the witness, suspect, defendant). Unfortunately, in both testing and in the law, arguments are sometimes problematic.

In previous sections of this paper key concepts and issues linking linguistics to legal issues were addressed through a case study, a review of key concepts, and a discussion of common myths, misunderstandings, and traps. Two key court constraints were also introduced contextualising the challenges when experts present evidence. In this section selected theoretical issues in assessment, specifically regarding validity, are examined for relevance to forensic linguistic cases involving NNSs. This will then be illustrated with one of a new NNS case.

Evidence argumentation in language assessment

Linguists/language assessment experts with research experience understand the importance of supporting research claims with evidence. However, there has been little discussion in the related fields on how to treat language proficiency issues in *individual* cases in forensic contexts. Some factors have been mentioned elsewhere in this paper. Usually there is a single subject. The perspective is usually retrospective related to communication in events in the past. The assessment must be as valid as possible for a very real, specific communication context. In such instances validity may need to override the high reliability of a standardised test or at least more valid assessment tasks can be used in conjunction with a standardised test. The expert's work is usually done under tight time conditions. If testifying, findings are presented in court where, in adversarial systems, they are vigorously attacked. The expert needs to be aware of possible gaps and ready to defend findings, and admit limitation, on the spot.

As already mentioned, judicial systems vary in their rules of evidence and what is acceptable as expert testimony. Also, whatever the rules, experts are expected to ground their work in the principles and research in the relevant fields. Fortunately, the language assessment field offers theoretical models that may be of use in forensic contexts. One of the several model builders working in this area is Robert Mislevy; he and his colleagues focus on applying evidentiary reasoning, evidence argumentation, to assessment (Mislevy, Steinberg and Almond 2003 and Mislevy 2006). He has also been influenced by the work of David Schum (1994) on foundational ideas on evidentiary reasoning drawing from research in law, philosophy, statistics, psychology, and other sources. Lyle Bachman has also used argumentation research/theory in his Assessment Use Argumentation (AUA) model for language assessment (Bachman 2005). In their 2007 LTRC (Language Testing Research Colloquium) presentation Chapelle, Enright and Jamieson explored the construct in interpretative argument, building on Kane's work in validation as well as Mislevy et al (2003). All cite the influence of Toulmin's 2003 work on practical reasoning.

While some readers are already well-grounded in evidence or assessment argumentation, for others, a brief introduction might be useful. The focus is on the work of Mislevy. McNamara and Roever present his assessment argumentation in their book, *Language Testing: The Social Dimension* (2006). It is aimed primarily at language assessment in educational, workplace, immigration/refugee contexts. Table 1 is from McNamara and Roever (p. 19) in which they summarise very clearly the assessment argument detailed by Mislevy, Steinberg and Almond (2003). The table is a useful graphic organiser to help identify both solid links as well as gaps in an argument. (This applies also if evaluating the claims of another expert.[8]) Across the first row are the basic

Table 1 The assessment argument

Evidence	→	Assessment argument	→	Claims about test takers
(Observations; assessment data)		(Relevance of data; value of observations as evidence)		(Inferences from observations)

From: McNamara and Roever (2006:19), Table 2.4: The assessment argument, drawing on Mislevy et al. (2003). Used with permission.

argument categories. In the second row are the realisations for each argumentation step. It is probably best that the user start with the last column filling in the claims to be made about the test taker, i.e., the suspect/accused. Claims are derived from legal issues in the case, e.g., 'It is highly likely that the defendant was not able to read and understand the law enforcement agent's summary of the interview'. Existing language assessment data, observations, and other language evidence are then entered in the first column. The middle column forces the expert to be clear about the quality of the links to the claims, the legal questions. A part of this process is addressing the issue of the validity of assessment instruments.

Gaps in argumentation may then trigger the need to (a) gather more background information, (b) reword the claim, (c) discard initial invalid assessment data, (d) do additional assessment, (e) examine in more detail test descriptors, (f) reconsider weight for validity and reliability, (g) do linguistic analyses of assessment data and other language evidence, and (h) consider still other options. At this point the expert would probably be in a relatively strong position to form an opinion on the claims related to the legal question. This may include determining that a claim cannot be supported. While some experts may not need graphic organisers for rigorous evidence argumentation, such a graphic (grounded in evidentiary reasoning) can be useful in presenting the language assessment data and findings to the court. This author also feels that if an evidence argumentation approach is used in a court, and if clearly presented, this may resonate with judges as it will build on their legal argumentation training.

While it is important to ground analyses and findings in language assessment theory, some in the language assessment field are concerned that paying attention only to 'technical measurement qualities' (i.e., reliability and validity) ignores the social consequences of testing. Thus, when developing and using tests the testing expert needs to consider the decisions that are to be made and the potential consequences of these decisions (Bachman 2005). This echoes the ethical concerns Shuy has raised (2006) that experts in forensic linguistics avoid an advocacy role and, more specifically, if moral issues are of concern, then stay away from a particular case.

Applying evidence/assessment argumentation in a forensic case

This section begins with an introduction to the case summary, followed by four general stages of argumentation, gradually increasing the theoretical validity of the assessment done and the grounds for assessing consistency/truthfulness in language performance.

Case summary

The case used in this section involves a non-native English speaking man who was charged with a felony crime. He was brought in twice for oral interviews with law enforcement officers (LEOs). The second appearance was to prepare him for a lie detector test that he had requested. However, the LEO said he believed the man and that no test would be needed. Nevertheless, an interview took place.

In both instances the LEOs took notes of the interview on a computer and then completed the document independently of the defendant. The defendant was then shown the document, asked to read it, and sign (initial) at the beginning and end of each paragraph as an indication that he understood and agreed with the contents. For the second interview, it took him no more than 10 minutes to read and initial each paragraph. There was no audio recording of either interview. The second interview document (language evidence) became the most critical as it was used as a confession to the crime.

Later when formally charged with the crime, he claimed he had not done it. When the document was discussed with him, he claimed that it did not represent what he thought he had said in the interview. When asked if he had actually read the document carefully, he admitted that he had not because it was long and he couldn't read English very well so he would take too long. When then asked why he had signed it, he replied that he had trusted the LEO who had said he believed him and that there was no need for a lie detector test.

Beginning the evidence/assessment argumentation

Readers may have many questions about the case, legal and otherwise. However, for the purposes of illustrating evidence or assessment argumentation, only one strand, related to interactive oral communication skills, will be developed here. Furthermore, the strategies will be only briefly presented as space does not permit a full, detailed argumentation report. (Reading comprehension was treated as a separate strand of argumentation although some relationship was shown as the document read and signed was a product of an oral interview.)

As anyone who has worked with evidence argumentation knows, reworking of the statements can seem endless in order to present the claims and evidence argumentation.[9] No doubt, readers will be able to suggest still further revisions. However, at some point the forensic linguist or assessment

expert needs to move forward with the best possible analyses and report at the time based on the court's timetable and the expert's own available time.

One of the general legal questions was 'How likely is it that the defendant was able to accurately understand the LEO and to express his story of the events in question?'. One of the linguistic questions was 'In the LEO interview what is the likely impact of the language proficiency of the defendant on being able to accurately understand and to accurately communicate details in a coherent manner?'.

One of the claims (from the defendant's perspective) was *'It is highly likely that the defendant had some difficulty communicating details accurately in the oral interview with the LEO'*.

Table 2 Beginning the argumentation

Evidence	→	**Assessment argument**	→	**Claims about test takers**
(Observations; assessment data)		(Relevance of data; value of observations as evidence)		(Inferences from observations)
Written statement of LEO Interview #2	**?**	NONE or No direct evidence	**?**	It is highly likely that the defendant had some difficulty communicating details accurately in the oral interview with the LEO. + It was highly likely that the defendant performed at a truthful level of oral proficiency.

First the claim was entered in the table (see Table 2). This established that evidence about the defendant's oral communication skills was required for this argumentation. When searching for language evidence of oral communication, the only documentation available was a written document ('statement') of the interview with the LEO, as written by the LEO. Therefore, legally there was no *direct* evidence as there was no audio recording. Even if a legal reporter/transcriber had been present, the transcript would not have been *direct* evidence for the many reasons a linguistics expert can give. Thus, the middle column of the table, assessment of the argument (relevance of data; value of observations as evidence) is filled in with 'No direct evidence' or 'None'. (This is marked by '?'.) In some legal settings the lack of an audio recording might be sufficient to throw out the evidence, but not in this case. Also because the defendant initialled each paragraph, it was assumed by the government that it was accurate. Forensic linguists have widely discussed the problems with written LEO summaries

even though the LEOs involved may sincerely think they have accurately represented the interview. An additional claim was also added by the defence in anticipation of a possible counter claim by the government: '*It was highly likely that the defendant performed at a truthful level of oral proficiency.*'

Strengthening validity

Collecting additional language evidence had to be done at a location very distant from the expert's home location. I made a cross-country flight and had one and a half days of actual working time on-site. Preliminary findings, while necessarily impressionistic, were also requested before the expert left the scene. As the communication setting was an interview, yet as there was no audio recording of the interview, I decided that an oral proficiency interview format/protocol was needed (a modified ACTFL-OPI). To check for the possibility of faking a lower than truthful performance, additional oral data were needed. Lacking a second examiner for the OPI, I added another oral assessment instrument, simulating an oral proficiency interview – a SPEAK test.[10] The two assessment instruments would probably meet the court requirement of being 'accepted' in the field and would provide external rating guidelines. Table 3 shows the additional evidence. It is recognised that with every testing instrument there are limitations. The assessment expert should be prepared to acknowledge the limitations if asked and to support the decision to use that instrument. (A general level of acceptance is marked by 'OK'.)

Going for stronger validity

While taking a lunch break after a morning of testing and meeting the legal personnel, I found I was still uncomfortable depending on the two formal external assessment instruments. The band descriptors did place him at similar levels (no correlation was claimed). Also, impressionistically there appeared to be consistency in performance (but the data were still to be analysed in more detail after I left the site). The defendant appeared to be somewhat effective in oral communication skills but not very accurate linguistically.[11] Still, the oral communications tasks lacked strong validity for assessing the likelihood that the defendant might or might not have had difficulty in the specific legal interviews with LEOs. I wanted a closer simulation of the LEO interviews, but it had to be much shorter.

Quickly I drafted tasks and a protocol and returned to meet the defendant again. Drawing on a work-related accident mentioned in the modified OPI, a short interview was conducted focusing on that accident.[12] Modelling the procedures used by the LEOs, I also drafted notes on the computer while interviewing the defendant and then edited it out of his presence. Later additional oral data were obtained (and recorded) in an oral feedback session on the written product of the interview.[13] While the simulation had not been previously tested, I felt it would contain more valid tasks than the other formal

Table 3 Adding language performance evidence

Evidence	→	Assessment argument	→	Claims about test takers
(Observations; assessment data)		(Relevance of data; value of observations as evidence)		(Inferences from observations)
Written statement of LEO Interview #2	**?**	NONE or No direct evidence More evidence needed	**?**	It is highly likely that the defendant had some difficulty communicating details accurately in the oral interview with the LEO.
+				
Audio recordings, scores and descriptors of – modified OPI – SPEAK		Formal instruments reasonably valid for oral interactive communication skills but only weakly valid for legal interview	**OK**	
		Scores when presented with range were informative	**OK**	
		Recorded language data useful for more detailed analysis of consistency in language performance	**OK**	+
		Band descriptors similar at a somewhat effective level and informative for fact finders *But the link still not strong*	**OK**	It was highly likely that the defendant performed at a truthful level of oral proficiency.

instruments and, thus, could be a potentially more valid source of language performance data.

In this third assessment his level of performance continued to be consistent with his performance on the other two instruments. With the simulated interview the apparent validity was strengthened. (This is marked with '✓'.) While I could not get inside the defendant's brain (no 100% certainty), I felt I could support the claim of 'highly likely'. Thus, as can be seen in Table 4, logically it could then be argued that there was some support for both claims.

Additional linguistic support

Even with my analyses of the language evidence from the three main language assessment sources, I felt research from other linguistic perspectives could also strengthen the claims. First, since none of the language evidence was direct evidence, whatever analyses I did had to be supported from various perspectives. The language proficiency data were selectively analysed from a second language development perspective. This further strengthened the claim of truthful performance.

Table 4 Going for stronger validity

Evidence	→	Assessment argument	→	Claims about test takers
(Observations; assessment data)	?	(Relevance of data; value of observations as evidence)		(Inferences from observations)
Written statement of LEO Interview #2		NONE or No direct evidence	✓	It is highly likely that the defendant had some difficulty communicating details accurately in the oral interview with the LEO.
+		More evidence needed		
Audio recordings, scores and descriptors of – modified OPI – SPEAK		Formal instruments reasonably valid for oral interactive communication skills but only weakly valid for legal interview	**OK**	
		Scores when presented with range were informative	**OK**	
+		All recorded language data useful for more detailed analysis of consistency in language performance	**OK**	+
		Band descriptors similar at a 'somewhat effective' level and informative for fact finders	✓	It was highly likely that the defendant performed at a truthful level of oral proficiency.
Short simulation of LEO interview		Validity increased as simulation tasks were similar to LEO interviews	✓	

Second, I also felt I could bring in other research that might strengthen the claim that the defendant probably had difficulty understanding the LEO in a 3-hour interview with no interpreter. I brought in linguistic research on oral interviews, on what each participant brings to an interview, on power/status of participants, and specifically on language assessment contexts and in law enforcement contexts. An additional area of theory and research was on contextualised v. decontextualised communication settings: an LEO interview is distant from the alleged criminal activity. Finally, I located some external information on his education and English language history. This all would further inform the fact finders (judge and jury) about the nature of the communications involved in the interviews.

Graphics were provided to illustrate the research. The prosecutor managed to persuade the judge that the graphic on the research on characteristics of interviews – on what the participants bring to an interview – was 'not relevant'

since 'it had nothing to do with' the specific interviews in question. The prosecutor's reasoning was that I was not actually present at the interviews! As no audio records had been made, I also provided information on how there has been a shift in some legal systems to require at least audio recordings of interviews and preferably video recordings where practical. This shift is occurring internationally in some countries and in some legal settings in the US. The prosecutor also successfully objected to a graphic on this shift on the grounds that it was not relevant to that particular legal system. I was not allowed to speak regarding these objections. Clearly more education of legal practitioners is needed regarding linguistics and language assessment.

Additional case notes

First, it should be noted that another strand of argumentation was also done in this case related to reading comprehension and the links to oral proficiency. Second, the evidence/assessment argumentation tables were not actually presented in my report nor in court. I used them as a tool (worksheets) to check my procedures and claims as I worked through the data collection and analyses. I knew I needed rigour, and even more so, since all the language evidence was indirect. (As I experiment more with the graphics, I may try to use the McNamara and Roever table format in court as long as it does not confuse the fact finders.) Finally, this first attempt to apply the tool of evidence/assessment argumentation allowed me to have greater confidence in my methodology and the findings I presented and forced me to be very clear and prepared for any limitations.

Concluding comments

Language is complex. Humans are complex. The layer of an additional language adds to the complexity of analyses of oral and written communication. Efforts to research and evaluate language proficiency and performance should involve multi-forms of methodology and assessment. Evidence/assessment argumentation provides one tool for language assessment experts to add rigour when developing procedures or evaluating links between language evidence and claims and the legal issues. A rigorous framing of argument derived from the language assessment field may help experts to:

- establish or evaluate the validity of links between language assessment data (and observations) and the claims required in specific forensic contexts
- consider the actual validity of the assessments used or proposed
- communicate with the court using an evidence argumentation perspective familiar in judicial settings
- ground arguments in respected theory and concepts from the language assessment community.

Notes

1. Paper presented at the ALTE Conference, Cambridge University, UK, April 9–12, 2008. Developed out of an exploratory presentation, July 2007, IAFL Conference, University of Washington, Seattle, Washington USA, 'Using Evidence Argumentation for Language Proficiency Assessment in Court' (van Naerssen 2007b).
2. In this case and in a later one, the first person singular is used as these are based on actual cases this author has worked on.
3. The interview protocol was intentionally modified from the normal ACTFL OPI in terms of length resulting from adding another role play and in terms of exploring the defendant's English language learning experience. Adding a role play increased the scope of language use, useful in detecting possible 'malingering' (in this case, pretending a lower than truthful language proficiency). Also this additional content on language learning experience is not allowed in an official oral proficiency interview, but I felt it might help me in evaluating consistency of language use in terms of language development. As this interview would be the only time I'd be able to meet with the defendant, I needed to collect this additional information and scope of performance at that time. I also needed a wider scope of the communications involved to be available as possible language evidence, thus, I felt might be more acceptable in court if it were collected within the oral proficiency interview. It is understood that the score was not an official ACTFL OPI score.
4. The defence counsel had also suggested instead, to reduce costs, that I simply sit in on his next meeting with the defendant. In this case I was new to forensic work, but my professional instincts told me to refuse. It was outside my formal assessment and analysis procedures. I also felt that it might bias my findings in several ways. I'm glad I did as later I learned that it very well could have caused my expert report to be impeached.
5. For example, a sentence repetition task to test a specific structure, might provide supplementary data if the task is carefully constructed (and includes distractor structures) and carefully administered.
6. If time and money permit, a second administration of the same oral test by a different examiner might be given. This may help address the issue of truthfulness of performance (faking or not faking a proficiency level).
7. Special thanks to the following for their input over several years: Lyle Bachman, Dan Douglas, Tim McNamara, Robert Mislevy, and Charles Stansfield. They have shared their ideas from their published research and given informal feedback on theoretical issues. They have looked at sections of drafts of several of my papers, brainstormed on challenges in particular cases, and shared their experiences and concerns in regard to testifying in cases involving NNSs. Any errors in interpretation remain of this author.
8. The graphic is used with permission of the authors (McNamara July 2008). However, this author is responsible for any errors in interpretation.
9. Chapelle et al. demonstrated in their 2007 LTRC presentation that the steps can be very time-consuming and possibly involve repeated looping back to revise assumptions, wording, and so on.
10. The SPEAK forms are retired forms of the ETS Test of Spoken English, a version of the Simulated Oral Proficiency Interview (SOPI) model, a tape-mediated test of speaking proficiency.

11. With scores of ACTFL-OPI Intermediate-Mid and SPEAK 40.
12. The simulated interview produced a reading product which was then used as a reading task, within the interview, similar to the production and use of written interview statement prepared by the LEOs. This was used in the reading comprehension argumentation.
13. The written material also provided data for another line of evidence argumentation regarding the defendant's reading comprehension skills. This line of argumentation is not presented in this paper.

References

Bachman, L (2005) Building and supporting a case for test use, *Language Assessment Quarterly* 2 (1), 1–34.

Chapelle, C, Enright, M, and Jamieson J (2007) (Where) is the construct in an interpretive argument? paper presented at the 2007 Language Testing Research Colloquium, Barcelona, Spain.

Cotterill, J (2002) *Language in the Legal Process*, Basingstoke, UK & New York: Palgrave Macmillan.

Coulthard, M and Johnson, A (2007) *An Introduction to Forensic Linguistics: Language in Evidence*, Milton Park, UK: Routledge.

Eades, D (2003) Participation of second language and second dialect speakers in the legal system, *Annual Review of Applied Linguistics* 23, 113–133.

Gibbons, J (2003) *Forensic Linguistics: An Introduction to Language in the Justice System*, Melbourne: Blackwell.

McNamara, T and Roever C (2006) *Language testing: The social dimension,* Oxford: Blackwell Publishers.

Mislevy, R J (2006) Cognitive psychology and educational assessment, *Educational Measurement* (4th edition), Phoenix, AZ: Greenwood.

Mislevy, R J, Steinberg, L S and Almond, R G (2003) On the Structure of Educational Assessments, *Measurement: Interdisciplinary Research and Perspectives* 1 (1), 3–62.

Office of Civil Rights (US Department of Education) (nd) Glossary, Programs for English language learners, http://ggsc.wnmu.edu/netc/synopsis/glossary.html

Ramirez, L (Ed.) (2007) *Cultural Issues in Criminal Defense,* 2nd ed., Huntington, NY: Juris Publishing.

Schum, D A (1994) *The evidential foundations of probabilistic reasoning,* New York: Wiley.

Shuy, R (2006) *Linguistics in the courtroom: A practical guide,* Oxford: OUP.

Toulmin, P (2003) *The uses of argument,* Cambridge: Cambridge University Press.

Underhill, N (1987) *Testing spoken language: A handbook for oral testing techniques,* Cambridge: Cambridge University Press, 88–103.

van Naerssen, M (2005) *Forensic linguistics: Can words help solve a crime?*, invited presentation, The Smithsonian Institution, Washington, D.C., September 21, 2005.

van Naerssen (2007a) Language proficiency and its relation to language evidence, in Ramirez, L (Ed.) *Cultural Issues in Criminal Defense* 2nd ed., Huntington, NY: Juris Publishing, 93–139.

van Naerssen (2007b) *Going from language proficiency findings to valid linguistic evidence*, paper presented at the 2007 IAFL Conference (International

Association of Forensic Linguists), University of Washington, Seattle, WA, USA, July 11–14, 2007.

van Naerssen (in press 2009) Introduction to forensic linguistics: Language content, language tools and communications with non-linguists, in Gueldry, M *Walk the talk: Integrating languages and cultures for the professions*, Vol 1 Cases, Lewiston, NY: Edwin Mellen Press.

4 A case of test impact: cheating on the College English Test in China

Dayong Huang and Mark Garner
University of Aberdeen, UK

Abstract

Cheating is one of the potential social impacts of high-stakes tests, and it has been investigated in the higher education systems in a number of countries. Little empirical research has been conducted, however, into the phenomenon in China. This paper reports the findings of a study of cheating on the national College English Test (CET) in China. The research examined the prevalence of cheating on the CET and the measures taken to counter cheating, the factors motivating students to cheat, and the consequences of cheating on the test. Data were gathered from a variety of sources using different techniques, and they show clearly that, whatever the actuality, cheating on the CET is *perceived* to be widespread, despite a number of countermeasures. The study identified five factors that motivate cheating, of which employers' use of the test results is by far the most significant. Cheating on the CET has induced consequences such as damaging institutional mission, threatening test validity and affecting test takers' beliefs in test fairness. The paper concludes that language tests like the CET have complex social impact that is as yet poorly understood.

Introduction

Research into cheating in higher education has shown that cheating is common, or at least is perceived to be so, in a variety of countries, such as Japan (Diekhoff, LaBeff, Shinohara and Yasukawa 1999), Russia (Lupton and Chapman 2002) and other former Soviet countries (Grimes 2004) and the United States (Whitley 1998). In China, the locus of the present study, Wan and Li (2006) found that more than 60% of college students cheated at times, and about 10% cheated in examinations. Over 82% of college students in Fu's study (2006) reported experiences of cheating.

The research reported here is the first stage of an ongoing project investigating the social impacts of the national College English Test (CET), which is used

to measure the English language proficiency of university students in China. This stage investigated the prevalence (and perceived prevalence) of cheating on the CET, the methods used, and its social and educational consequences.

The CET was introduced by the Ministry of Education in 1987; by 2006 the test population had reached 12 million (Jin 2007). These numbers continue to grow, and with them the influence of the CET on the lives and academic pursuits of university students. Although it was designed purely to assess linguistic proficiency, the CET has also been used to serve other ends. Several years ago, some universities began to use CET certificates as a hurdle requirement for degrees. Negative public reactions led to the reduction of this practice, but many employers use CET results as a selection criterion for graduates, even for positions in which knowledge of English is not necessary. The CET thus exerts an influence on more than simply language learning. Although the social impact of language tests has been addressed by some scholars (e.g. Kunnan 2005, McNamara and Roever 2006, Shohamy 2001), there is a lack of empirical studies. The present study is a contribution to filling this gap, and, despite its limited scope, the findings suggest that the CET is a rich field for applied research into not only cheating, but many other aspects of the educational and social impacts of language testing.

The study addressed four key questions:

1. How prevalent is cheating on the CET?
2. What are the major factors influencing students to cheat?
3. What counter-measures are adopted to try to prevent cheating?
4. What are the perceived social and personal consequences of cheating?

The primary data were gathered in 20 semi-structured, in-depth interviews conducted in 2007. The sample comprised 12 undergraduates and three highly experienced and senior university administrators at one Chinese university, and five Chinese postgraduates currently enrolled at a British university. The interviews were in Mandarin, but excerpts have been translated into English for this paper.

The interview data were supplemented from other sources:

- the lead author's own experience of administering the CET for more than 10 years as a faculty member in a university in China
- participant observation of the administration of the CET at one university on 23 June 2007
- an examination of reports in the media
- monitoring of discussions of cheating in a Chinese online forum
- examination of official documents.

Two kinds of cheating can be identified: cheating on one's own behalf, and assisting others to cheat (by, for example, acting as surrogates – called 'gunmen' in China – or collaborating in discovering and providing answers).

The primary focus of this research is on the former group, although the latter is touched upon in places.

How prevalent is cheating?

It is very difficult to gain a reliable picture of the extent of cheating on the CET. Statistics refer to the numbers of candidates convicted, and these suggest a relatively small but persistent problem. For example, in June 2004, 138 test takers in 10 universities in Shan'xi Province were convicted of cheating (Xu and Chen 2004). In 2005 and 2006, 861 and 1,973 convictions respectively were reported in Hubei Province (The Educational Testing Centre of Hubei Province 2006). In the latter year, 109 convictions were recorded in Beijing: 97 involved test takers, and 12 involved confederates who transmitted answers to candidates during the test (Du 2006). The figures, however, understate the actual frequency of cheating, as witness the fact that, on two occasions when the first author helped to administer the CET, 10 students were detected cheating, yet not reported to the public. The reports of interviewees bear this out:

> Too few. It is just a small part. There are lots of cheaters. (Student 1)
> Those caught are just a small part of those who cheated on the CET. (Student 2)
> Those caught were just the small part of all who cheated on the test. (Administrator 1)

All students interviewed cited examples of successful cheaters known to them personally. Moreover, if even a small *proportion* of students are convicted, given that there were 12 million test takers in tests in 2006 (Jin 2007), the actual numbers would have been considerable.

Statements by senior officials are liable to minimise or overstate the problem according to political considerations, whilst certain high-profile cases reported in the media may give a distorted impression of the level of cheating involved. For example, a university officer who was in charge of storing CET examination papers leaked the test contents to a teacher in a coaching school for commercial purposes. Both were sentenced to three years in jail. In 2004, another two separate court cases resulted in prison sentences for several proctors and officers who sold test items to students for large sums of money (Wang and Yang 2006). These and similar cases have received extensive media coverage.

Why do students cheat?

The interview data indicated five significant factors associated with cheating:

1. motivation
2. opportunity

3. peer pressure
4. failure of enforcement
5. facilitating conditions

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor (Campbell 1976:49).

Suen and Yu (2006), in their study of cheating on the Chinese Civil Service Exam (called *keju*), concluded that as long as there is competition in society and stakes attached to the test, cheating is unavoidable. Security and punitive measures will not solve the problem.

Motivation

The overwhelming motivation to cheat on the CET is the fact that employers use the results in selection for jobs. Despite China's economic boom, rapidly expanding enrolment in higher education (enrolments were in the region of six million in 2008, up 5% from the previous year) has resulted in deteriorating employment prospects. In 2007, more than 530,000 applicants competed for 12,700 government jobs: an average of 42 applicants for each position, and about 20% of graduates were unemployed in 2007 (Yan 2008). Interview evidence is that the rate of unemployed graduates from certain universities may be as high as 40%. In this extremely competitive job market, it is crucial for university students to obtain a CET certificate. As Student 2 said, 'If it weren't for a job, no one would cheat'; as Administrator 1 commented, 'test results may have life-long influence on [students]'. Deteriorating prospects for graduates have led students to try to obtain as many certificates as possible to present to the employers in their résumés. The CET certificate has become the most important one because employers usually use it as a 'door keeper' in selection:

> Employers . . . put their requirements on the door of the interview room. If you did not pass CET-4, you could not get in the door. (Student 5)
> The selection manager said before the interview: 'no CET-6 qualification, no interview'. (Student 6)
> . . . there is a general rule: if you passed CET-4, the employers select you; if you passed CET-6, you select the employer; if you did not pass CET, no employer will select you. (Administrator 3)

Family expectations increase the motivation:

> The parents have contributed so much for our education . . . Can you imagine how great the pressure would be? . . . some students risk danger in desperation. (Student 5)

The motivation to cheat increases with every unsuccessful attempt at the CET:

> In the [final] year, all despaired of passing the CET; they had to buy answers. (Student 4)

In an online forum on the topic (Online commentator A 2006), several participants seemed to believe that cheating is an accepted part of the CET process, e.g.:

> This is the fourth time for me to take CET. I bought answers and earphones . . . When I came out of the test site, I saw many students celebrating their success in cheating; students caught cheating were anxious to seek connections or bribe the teachers.

Opportunity

Motivation alone will not occasion wrong-doing: there must also be opportunity. Students believe that the test format provides this. It is easier to cheat on the so-called 'objective' (multiple-choice questions) elements of the test: for instance, by buying the answers in advance or having them transmitted from outside during the test. It is more difficult, although not impossible, to cheat on 'subjective' questions requiring free composition. The test format has been changed in recent years. Previously, the pass mark was 60. The maximum score on multiple-choice questions was 85. The subjective item (composition) was worth 15 marks, but was weighted: a score of 0 resulted in failure on the whole test; a score of 1 to 5 meant a deduction of 6 marks from the total score. This device was, however, ineffectual: it takes very little to achieve a mark of at least 1 or 2 in writing, so a student who scored around the mid-sixties on the multiple-choice questions would still pass the CET (e.g. $65 + 1 - 6 = 60$). In 2007, objective items were reduced to 75 and there was no minimum score for the subjective element. The opportunity to cheat is still present, however:

> Even though the test has been reformed . . . it is quite possible to pass by finishing only these objective items and ignoring the subjective parts. This gives students the chance of cheating. (Student 3)

Peer pressure

McCabe (1997), McCabe and Trevino (1993), McCabe and Drinan (1999) and McCabe, Trevino and Butterfield (1999, 2001, 2002) studied the correlation between college cheating and both individual and contextual factors. They

found social factors – peers' behaviour (McCabe and Trevino 1993) and peer disapproval (McCabe and Drinan 1999) – are the most influential because:

> academic dishonesty not only is learned from observing the behaviors of peers, but peers' behavior provides a kind of normative support for cheating . . . The cheating may come to be viewed as an acceptable way of getting and staying ahead (McCabe et al 2001:222).

These factors are enhanced within Chinese culture by the important role of *guanxi* (social connections). When facing a penalty, students try to utilise their connections to have it reduced or withdrawn. When they are successful, as they sometimes are, this strengthens other students' belief that they can cheat with impunity. A contributor to an online discussion noted sardonically:

> Would-be Party members who have really set a 'Pioneering Model Example' for other students: when they were caught cheating, they were not punished according to the Party Regulations. They were OK after writing a paper of self-criticism (Online commentator B 2007).

Furthermore, there is a general perception of misconduct in almost every part of society: corruption among government officials; plagiarism by university professors; sharp practices in business. It is not surprising, then, that in an online survey conducted by China Central Television (2005), over half of 8,454 respondents 'would not mind' if another person cheated on a test (provided it does not affect you). Over a quarter found it 'difficult to say'.

Failure of enforcement

In the online survey mentioned above, 32.4 % of the respondents expressed the opinion that cheating can be engaged in with impunity. There is a relatively low probability of being caught and, if caught, of being punished. In the instances cited earlier, the relatively severe custodial sentences were given to administrators and teachers; students are typically treated more leniently, and security regulations and invigilation procedures are at best inconsistently enforced.

Current regulations stipulate two proctors for each standard test room of 30 students. The large numbers taking the test mean that there is always a shortage of independent proctors, so most universities train teaching staff to act in this capacity. Since the teachers know the students personally, however, they are less inclined to enforce the rules. Moreover, convictions require hard evidence, properly recorded, of any alleged misdemeanour. Consequently, the emphasis in training sessions is on prevention rather than conviction. The temptation is to save trouble by looking the other way:

> Generally, cheating students will not be caught . . . the administering teachers are not very strict and fussy. (Student 2)

Similar views were voiced in almost every interview.

Even if a student is caught, the penalties are often light and inconsistently applied. Universities have different forms of punishment. Some give a warning or document a penalty in the student's record. During the interviews, Student 8 told how she and a fellow student were caught; they were given a recorded warning and their degree certificates were withheld. Two confederates, however, were not punished. Student 6 knew someone who was convicted, but:

> She was not really punished. Maybe the university wanted to give her an opportunity to correct her mistake.

Administrator 2 admitted that the universities are in an invidious position:

> The university does not give severer punishments because punishments have great influence on their future. The university only punishes those serious cheating students and leaves some others unpunished.

Fear of damaging the university's reputation is probably also a factor as shown by a notice issued by the university:

> If there is any serious problem on the test, the university's reputation will be greatly affected. All schools and related departments should organize before-test education of the test takers to try to eliminate . . . cheating.

Facilitating conditions

In the words of Administrator 2:

> Cheating on the CET is guided by the law of market economy: where there is demand, there is supply.

Modern communications technology facilitates supply. One method is for several collaborators to complete different parts of the test and send the answers by mobile phone to confederates outside, who then supply them to clients. Another is to transmit photographs of the test papers as soon as they are distributed. The following reputedly eyewitness account (mlbswy0828 2006) gives an idea of the sophisticated organisation involved:

> A group of [students] entered the Internet bar and logged on. A female rushed in . . . 'Hurry, these are the answers to test paper A'. My God, they were a gang of CET answer providers! . . . These people divide themselves

> into three small groups, 3 persons each group. One person is sending answers on the QQ to clients; another is making calls: Hello! If you have Test paper A, just hold on. If not, hang up. Paper B's answers will come later. The third person then rushed forward: 'Here are answers to Paper B' . . . They even read an essay sentence by sentence to the phone . . . They divided the tasks clearly, cooperated very well and operated skilfully.

Combating cheating

Various measures have been implemented by the CET committee to combat cheating, and are used in various combinations by different universities.

Moral education

One is moral education, which is an important part of the curriculum at all levels in China. Recent CET regulations require all test takers to sign a pledge. It is doubtful, however, that this is effective: in the online survey (Huang 2008), 71.5% of the 2,106 respondents agreed 'it is a formality' and 88.8% felt 'honesty is not guaranteed by just signing a promise'. Contributors to an online discussion (Online commentator B 2007) agreed:

> It is no use signing an honesty promise in examinations. We have regulations that are not implemented and we have laws that are not put in effect. All these are like a blank paper.
>
> What is the use of signing a pledge? . . . This is like those Party members who pledged under the national flag. Some of them would be loyal officials and some will be corrupt.

Some universities actively promote honesty. For example, in some, classes with a clean record take exams without invigilation: teachers distribute the test papers, leave the room, and return to collect them at the end. If cheating is found to have occurred, all are punished. One university initiated the practice of rewarding a 'Zero Cheating Class' with 100 yuan. If all pass the test, the class receives a further 500 yuan (Zhang and Lei 2002). An administrator is quoted as saying:

> It is not very effective to . . . issue official notices and propaganda. We . . . take another approach: to foster the awareness among students of staying away from cheating and to make them morally adapted to a no-cheating test.

Opinions concerning the experiment vary. One student commented:

> Being honest in test is the basic requirement for students. It is ridiculous to award a 'zero cheating' class with money (Zhang and Lei 2002).

In a more artistic vein, another student wrote (Wu 2008):

> In the final exams . . . there was a flag above the blackboard: 'Zero Cheating class of the School of Information Science'. This is the new idea figured out by our school to prevent cheating on exams. The flag is a show to both the invigilators and the students in an attempt to stop the endless phenomenon of cheating. However, the flag was not capable of stopping the 'cheating wind' coming with the spring wind. What was happening in the test room can be described with two informal poetic lines:
>
> The red flag is swinging on the stage,
> The cheating notes are flying under the stage.

Punishment

A more conventional means of dissuading cheats is punishment. According to the CET administration manual, if a student is found cheating on the test, he or she should be stopped and taken out of the test room. A zero mark is recorded, and the case is recorded and reported in writing to the CET committee. Any further action is left to the university, and consequently, as noted above, penalties differ from university to university. They commonly involve a public notice on the campus giving the names of those who are caught cheating and the penalty meted out.

The severest penalty is summary dismissal, but universities are cautious about applying it, as some students dismissed for cheating have successfully sued their universities, which have been forced to rescind the penalty. Many avoid litigation by stopping short of dismissal, whilst entering the conviction on a student's record. Any further punishment, such as refusing the degree, is decided on the circumstances of each case, which reinforces the perception of inconsistency. Unsurprisingly, there are many public calls for a universally applicable set of penalties. The Ministry of Education, however, maintains that appropriate codes of punishment already exist in *Penalties for the Violation in National Examinations* and in university regulations and honesty record systems (Shi 2006). Despite such pronouncements, it is clear that punishments are widely perceived as neither efficacious nor consistent.

As in all areas of human activity, the punishment for cheating must be *proportionate*. On the one hand, they must not be so light as to lose any deterrent effect; on the other hand, severe penalties may have grossly disproportionate consequences for the culprit, and are increasingly being challenged in the courts. In the absence of extensive research findings relating to the CET, the educational, moral, and behavioural outcomes of the many different types and levels of penalties are unknown.

Prevention

Security is constantly being enhanced, through both improvements in the way the test is administered and increasingly sophisticated technology to detect and counteract attempts at communication between collaborators. National regulations for the administration of the CET are published in the *CET Administration Manual*, and additional measures appropriate to local conditions may be taken. Test papers must be securely locked away until needed; delivered in sealed bags to the site by at least two officers; handed over to proctors half an hour before the start of the test; and opened in the room 15 minutes later. Proctors may not leave the building, which is guarded.

The CET regulations prohibit anyone other than university students from taking the test. Test takers are photographed at the point of registration, and in some universities submit to a fingerprint identification system. When entering the test room, test takers sign in, and are searched for electronic equipment such as mobile phones. They are allowed to leave the room, accompanied by a proctor, only in exceptional circumstances. In no circumstances may they leave the building before the end of the test (see Pictures 1–2).

The growing use of communications technology by cheats has brought about increased emphasis on measures both to detect the presence of telecommunication devices and to block transmissions around test sites. Signal-

Picture 1 Guards at the entrance of a test site (taken by the first author during the test)

Picture 2 Identification documents to be checked by the proctor and identification sheet to be signed by test takers as they enter the test room (taken by the first author during the test)

blocking equipment is placed on each floor of the test site. Each room is also fitted with a detector; a radio signal will set off the alarm (Picture 3). Some universities employ patrol cars equipped with probes to monitor all civil radio frequencies and to pinpoint the source of signals. Nonetheless, in the opinion of Student 1:

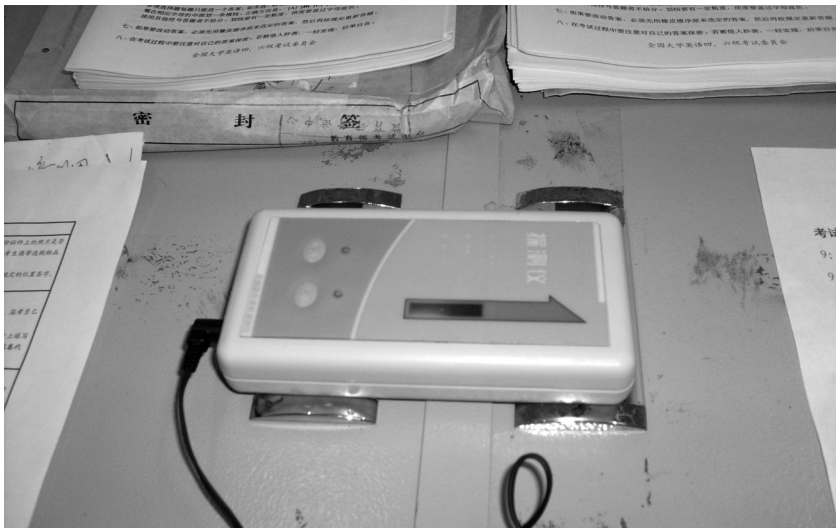
> Generally speaking, the cheating-detecting devices are two or three years behind the cheating devices.

A technological arms-race between cheats and administrators is likely to continue indefinitely.

Detection through analysis of answers

This approach to detecting cheating has been studied in the name of 'authorship attribution techniques' in the literature of forensic linguistics (see Brooks, this volume). Studies show that different limitations are inherent in these techniques, which means that inaccuracies can arise from their application; nonetheless, they have been used to detect cheating on the CET. One of the authors (Huang) had personal experience of the results of this method several years ago, when he and a fellow-invigilator were reprimanded for not preventing cheating, which had been detected by post-test analysis. No other

Picture 3 A signal detection device in a test room (taken by the first author during the test)



documented or anecdotal evidence for the use of this method could be found, however. It is possible that the reduction of multiple-choice questions, mentioned above, has lessened the value of analysis for detecting cheating.

The use of a combination of counter-measures by one university was described by Administrator 1:

> We have been taking all measures to reduce or curb cheating. We educate students to be honest. Students are required to sign up to a code of honesty before the examinations. The invigilators are required to alert or warn students when they want to cheat. Test irregularities are timely handled when they occur . . . In the future, the government may record cheating as bad credit records, adding to individuals' own honesty system. If this is implemented, cheating would be very risky for an individual . . . Of course, cheating should mainly be prevented through moral education and technology. We educate students not to think of cheating and we apply technology to counter cheating. We should also create a positive learning environment and motivate students to work hard in learning. Cheating problems should be solved with the combination of these measures.

The limitations of counter-measures

It is clear from the evidence given earlier that the extensive and multi-faceted battle against cheating is not as effective as administrators would like. Pursuing the battle too forcefully can lead to unintended and undesirable

consequences for the university. A rigorous programme of policing and detection causes additional stress for the honest majority of test takers. Casting faculty and students into adversarial roles of police and criminals, with the implication that students have no ethical code and will cheat whenever they can, can be counter-productive. It can mar the student–teacher relationship and hence the whole educational process; it can also encourage some students to live up to the role in which they have been cast. This is an important area for applied educational research.

The impact of cheating

Cheating in schools and universities might lead to four consequences (Passow, Mayhew, Finelli, Harding and Carpenter 2006:678):

1. Undermining institutional missions that include preparation for citizenship and service to society, each of which has a moral dimension.
2. Invalidating measures of student learning and grading equity.
3. Damaging student and faculty morale, the reputation of the institution, and public confidence in higher education.
4. Increasing the likelihood of dishonest acts both outside the classroom and after graduation.

The present study found evidence of just such consequences in relation to the CET, together with another, not mentioned by Passow et al:

5. Physical injury.

Undermining institutional missions

University students are regarded as the backbone of the society in China. The universities are seen as a holy palace where the university students are educated to be elites of the society. Academic cheating by the university students is not only damaging the image of the university but also undermining its institutional mission. As one kind of academic dishonesty, cheating on the CET is dragging many university students into unethical conduct which may continue after they enter society, as suggested by some studies (Carpenter, Harding, Finelli and Passow 2004, Lawson 2004, Sims 1993). Administrator 2 articulated his concerns:

> Our universities are supposed to turn our young generation into high-quality citizens with morale and ethics, not skilful cheats. If we cannot solve the cheating problems in higher education, we bring not only shame to ourselves but also disaster to our nation.

Threatening test validity

As Suen and Yu (2006:56) observe, 'When cheating occurs, the lack of validity is incontrovertible'. The scores obtained by CET cheats, for example, are not true indicators of their ability in English. Informants cited examples of cheats who had scored very highly, and even topped the class. The effects are not limited to the offenders themselves, however. Student 2 (who did not cheat) complained that his performance on the listening comprehension – transmitted by radio – was affected by the activities of others:

> I could not listen clearly when doing the listening comprehension. It was too noisy. I did not know why. After the test, I asked and found that my radio signal was affected by the signals the cheating students were receiving.

Many similar cases are reported online (Du 2006).

Damaging belief in test fairness

Cheating not only brings about factual unfairness in testing but also damages the test takers' trust:

> I felt unfairly done by. I studied so long and did not pass. [The cheats] passed very easily by paying several hundred yuan . . . But nothing can be done. This is the reality. Many people were affected, not only me. (Student 2)

> It is unfair! We worked hard and passed with a low score. [The cheats] scored very high, even above 500 points. They could take advantage of the high scores in job hunting. (Student 13)

Increasing dishonest activity

As has been shown, cheating on the CET implicates not only the test takers. Almost always, other students – sometimes many others – are involved in each occurrence. On occasion, the motivating factor is simply the desire to help a friend, as Student 8 confessed:

> I did it for my friend. He asked me to send him a text during the [CET]. . . . I stayed outside and received answers and transferred the answers to him with my mobile phone. He was caught inside and the proctors found my number from his phone . . . I did not know the serious consequences of doing this. I just agreed to help a friend without any thought.

The practice of surrogate test takers is well known. Reports in the media suggest that some students have become, in effect, professionals in surrogacy.

One student sat and passed three national postgraduate entrance examinations and two national judiciary examinations. His efforts earned him in all 60,000 yuan (about £4,500). He has also taken the CET several times, and is known for his lavish spending (Guo and Li 2004). Another student in the same university successfully sat the CET eight times. His charges were approximately £100 for CET-4 and £180 for the CET-6 (Guo and Li 2004). There are reports that such 'professionals' are sometimes aided, either actively or passively, by proctors – who might also be paid (Niu and Woff, no date).

There is a well-organised trade in cheating devices, conducted by students or outsiders:

> Advertisements are everywhere. Online buying is not reliable. It is more reliable to buy through the introduction of the classmates. Both know each other and the buyer can get the money back if there is any problem. (Student 1)

> There are agents selling the devices. Some students do this too. (Student 3)

Physical injury

The minute size of modern communication devices makes them an attractive proposition for cheats, but also presents a serious threat of injury:

> The earphone is very tiny, smaller than a particle of rice. It is very easy to slip into the deeper part of the ear and injure the eardrum. And it is very difficult to take it out by ourselves. (Student 2)

There are widespread reports of such injuries, which sometimes require delicate surgery. On a single test day, one doctor in Hubei operated on six students to remove earphones (Southcn 2006), and a hospital in Guangzhou reported 30 such operations (Huang, Liu and Huang 2007). It has been known for an interphone hidden in a student's abdomen to explode and cause internal bleeding (Guan and Ma 2006).

Conclusion

It is obvious, even from the present small study, that cheating is not a simple phenomenon. There are multifarious motives for cheating and many different methods by which it is accomplished. It is still quite unclear what the appropriate response – or, better, range of responses – should be. However, it is clear that the use of high-stakes language tests like the CET has an impact beyond the immediate learning and teaching situation. It is also clear that this impact in turn influences the qualities of the test (such as validity) and

produces more profound effects on test takers, on the education system and on the society. Thus, there are circular and chain consequences of using high-stakes language tests. This reminds us of one of the most memorable quotes in language testing: 'Tests should be labeled just like dangerous drugs. Use with care!' (Spolsky 1981:20).

The CET and similar high-stakes tests are both 'social practice and social product' (Filer 2000). The present study not only provides empirical evidence of the educational and social impacts of language testing, but also indicates the need to study those impacts from wider contexts (Kunnan 2005) and from interdisciplinary perspectives (Spolsky 1995:3) by taking the 'socio-cultural aspects of assessment' (Gipps 1999) into consideration.

References

Brooks, R L (2009) In your own words, please: Using authorship attribution to identify cheating on translation tests, in Taylor, L and Weir, C J (Eds) *Language Testing Matters: Investigating the wider social and educational impact of language assessment*, Cambridge: UCLES/Cambridge University Press, 77–102.

Campbell, D T (1976) *Assessing the impact of planned social change*, Paper #8 Occasional Paper, Accessed 15/02/2006 at <https://www.wmich.edu/evalctr/pubs/ops/ops08.pdf>.

Carpenter, D D, Harding, T S, Finelli, C J and Passow, H J (2004) Does academic dishonesty relate to unethical behavior in professional practice? An exploratory study, *Science and Engineering Ethics* 10 (2), 311–324.

China Central Television (CCTV) (5/2/2005) *Analyzing the cheating phenomena on national education examinations*, Accessed 20/02/2008 at <http://news.sohu.com/20050205/n224251328.shtml>.

Diekhoff, G M, LaBeff, E E, Shinohara, K, and Yasukawa, H (1999) College cheating in Japan and the United States, *Research in Higher Education*, 40 (3), 343–353.

Du, D (25/12/2006) *CET leaking again? 97 cheats caught in Beijing*, Accessed 20/02/2008 at <http://edu.people.com.cn/GB/5211642.html>.

Filer, A (2000) Introduction, in Filer A (Ed.) *Assessment: Social Practice and Social Product*, London: RoutledgeFalmer, 1–5.

Fu, X L (2006) College students cheating on examinations, *Science and Education*, Feb (second half), 34–35.

Gipps, C (1999) Social-cultural aspects of assessment, *Review of Research in Education* 24, 355–392.

Grimes, P W (2004) Dishonesty in academics and business: A cross-cultural evaluation of student attitudes, *Journal of Business Ethics* 49 (3), 273–290.

Guan, X F and Ma L (20/6/2006) *High-tech products used for cheating in examination*, 25/03/2008 at <www.chinadaily.com.cn/china/2006-6/20/content_620957.htm>.

Guo, G Z and Li, R X (23/12/2004) Investigating the chains of profiting from the cheating on the CET, *Oriental Outlook* 27, Accessed 15/07/2007 at <http://www.gmw.cn/content/2004-12/23/content_102355.htm>.

Huang, C (21/01/ 2008) Survey: 55.2% Believe widespread dishonesty on the CET, Accessed 20/02/2008 at <http://edu.sina.com.cn/cet/2008-01-21/0955118945.shtml>.

Huang, Z, Liu, Y and Huang, W (23/12/2007) *CET administered yesterday, much cheating still reported*, Accessed 25/03/2008 at <http://news.xinhuanet.com/edu/2007-12/23/content_7298535.htm>.

Jin, Y (2007) *Powerful tests, powerless test designers? New Challenges Facing the College English Test,* presentation at the 5th International Conference on ELT in China & the 1st Congress of Chinese Applied Linguistics, Beijing, May 16–21, 2007.

Kunnan, A J (2005) Language assessment from a wider context, in Hinkel, E (Ed.) *Handbook of Research in Second Language Teaching and Learning*, Mahwah, NJ, USA: Lawrence Erlbaum Associates, Incorporated.

Lawson, R A (2004) Is classroom cheating related to business students' propensity to cheat in the 'real world?', *Journal of Business Ethics* 49 (2), 189–199.

Lupton, R A and Chapman K J (2002) Russian and American college students' attitudes, perceptions, and tendencies towards cheating*, Educational Research* 44 (1), 17–27.

McCabe, D L (1997) Classroom cheating among natural science and engineering majors, *Science and Engineering Ethics* 3, 433–445.

McCabe, D L and Drinan, P (1999) Toward a culture of academic integrity, *Chronicle of Higher Education* 46 (8), B7.

McCabe, D L and Trevino, L K (1993) Academic dishonesty: Honor codes and other contextual influences, *Journal of Higher Education* 64, 522–538.

McCabe, D L, Trevino, L K and Butterfield, K D (1999) Academic integrity in honor code and non-honor code environments: A qualitative investigation, *Journal of Higher Education* 70 (2), 211–234.

McCabe, D L, Trevino, L K, and Butterfield, K D (2001) Cheating in academic institutions: A decade of research, *Ethics & Behavior* 11 (3), 219–232.

McCabe, D L, Trevino, L K, and Butterfield, K D (2002) Honor codes and other contextual influences on academic integrity: A replication and extension to modified honor codes settings, *Research in Higher Education* 43 (3), 357–378.

McNamara, T and Roever, C (2006) Language Testing: The Social Dimension, *Language Learning* 56, Supplement 2, Michigan: Blackwell Publishing.

Mlbswy0828 (2/3/2006) An astonishing scene on the day of CET administration, Sina BBS, Accessed 25/03/2008 at <http://edu.sina.com.cn/bbs/2006/0303/183802117.html>.

Niu, Q and Woff, M (year unknown) *Chinese University Diploma: Can its international image be improved?* Accessed 25/01/2008 at <http://www.usingenglish.com/esl-in-china/diploma.pdf>.

Online commentator A (18/6/2006) Topic: *Last time to take the CET, What should I who do not want to cheat do?* Accessed 20/02/2008 at <http://comment.education.163.com/reply/post.jsp?type=null&board=education2_bbs&threadid=2JNSA8BK00291OUM&showdistrict=&pagex=3>.

Online commentator B (24/12/2007) Controversies over signing honesty pledges by CET test takers, XinhuaNet BBS, Accessed 25/03/2008 at <http://view.home.news.cn/comment?url=http://news.xinhuanet.com/edu/2007-12/23/content_7297642.htm>.

Passow, H J, Mayhew, M J, Finelli, C J, Harding, T S and Carpenter, D D (2006) Factors influencing engineering students' decisions to cheat by type of assessment, *Research in Higher Education* 47 (6), 643–684.

Shi, X (Ed.) (26/12/2006) *Ministry of Education: No leaking and large-scale cheating on the CET*, Accessed 25/03/2008 at <http://edu.people.com.cn/GB/5214642.html>.

Shohamy, E (2001) *The Power of Tests: A Critical Perspective on the Uses of Language Tests*, England: Pearson Education Limited.
Sims, R L (1993) The relationship between academic dishonesty and unethical business practices, *Journal of Education for Business* 68 (4), 207–211.
Southcn (20/6/2006) *Mini earphone in the ears of CET test takers*, Accessed 25/03/2008 at <http://education.163.com/06/0620/16/2K2SBB8900291OSH.html>.
Spolsky, B (1981) Some ethical questions about language testing, in Klein-Braley, C and Stevenson, D (Eds) *Practice and Problems in Language Testing*, Frankfurt: Verlag Peter D Lang, 5–30.
Spolsky, B (1995) *Measured Words*, Oxford: Oxford University Press.
Suen, H K and Yu, L (2006) Chronic consequences of high-stakes testing? Lessons from the Chinese Civil Service Exam, *Comparative Education Review* 50 (1), 46–65.
The Educational Testing Centre of Hubei Province (20/6/2006) *General Reports on the CET administration in our province*, Accessed 20/02/2008 at <http://hbea.hubei.gov.cn/Article_Show.asp?ArticleID=1046>.
Wan, Y K and Li, H T (2006) A study on college students cheating in examinations, *Journal of Yuxi Teachers College* 22 (3), 86–90.
Wang, L and Yang Y (7/11/2006) *The underground industry of cheating on tests*, Accessed 20/02/2008 at <http://news.people.com.cn/GB/37454/37462/5007062.html>.
Whitley, B E (1998) Factors associated with cheating among college students, *Research in Higher Education* 39 (3), 235–274.
Wu, C Y (27/1/2008) '*Zero-cheating Class': Students' shame or education's shame?* Sohu Blog Accessed 25/03/2008 at<http://discoverly0-0.blog.sohu.com/ >.
Xu, H and Chen, G (24/6/2004) *CET in Shanxi: 138 punished for cheating*, Accessed 20/02/2008 at <http://zqb.cyol.com/content/2004-06/24/content_894887.htm>.
Yan, L (16/1/2008) *Increase of college enrollment in China to hit decade low*, Accessed 20/02/2008 at <http://news.xinhuanet.com/english/2008-01/16/content_7433395.htm>.
Zhang, S W and Lei, Y (2/7/2002) *Questioning 'Zero-cheating Award'*, Accessed 25/03/2008 at <http://news.xinhuanet.com/st/2002-07/02/content_466239.htm>.

5 In your own words, please: using authorship attribution to identify cheating on translation tests

Rachel L Brooks
Federal Bureau of Investigation

Abstract

Cheating on a Federal Bureau of Investigation (FBI) translation test results in serious consequences for both the agency and the examinee; therefore, authorship attribution techniques have been trialled to determine whether they can provide a replicable, valid procedure for detecting cheating. The research questions examined whether the techniques of lexical overlap, type/token ratio, richness score, hapax percentage, collocations, and shared hapaxes effectively support a claim of cheating by plagiarism in translation tests and the amount of similarity between two translations required for them to be considered plagiarised. Two sets of data were retrieved from FBI translation tests in Arabic and Urdu, each representing a known case of cheating. Texts in question were compared to each other and to a corpus of 40 other test responses. Analyses employed were borrowed from research on authorship attribution techniques used in research and in courts. Results revealed that though many attribution techniques did not indicate cheating from the two suspect texts, the collocations and shared hapaxes appeared effective to that end. Suggestions are made for a set of baseline statistics to determine the minimum evidence needed for cheating in translation tests.

Introduction

Cheating is on the rise in the United States. Whereas in the 1940s, about 20% of students at colleges claimed to have cheated during academia, today the number has increased to between 75% and 98% (ETS 1999). Cheating is not limited to high school; it has been found in such high-stakes tests as those used to qualify translators at the Federal Bureau of Investigation (FBI). In this case, like so many others, the consequences of unreliable test scores are quite serious, potentially resulting in faulty translations that could put FBI

agents' lives in jeopardy. Even though instances of cheating on FBI translation tests are few, the need to establish a reliable methodology to determine whether or not cheating has occurred remains critical. Authorship attribution techniques employed in Forensic Linguistics may be useful for this purpose.

The goal of this study was to determine whether or not instances of suspected cheating on FBI translation tests could have been determined by *post hoc* analyses using authorship attribution techniques. Such techniques would allow the investigator to examine the amount and types of variation that were typically present among translation test responses and isolate instances where two different examinees' responses are abnormally similar, causing them to be statistical outliers from the rest of the corpus. The techniques of lexical overlap, type/token ratio, richness score, hapax percentage (the ratio of words that appear only once to words appearing multiple times in a text), collocations, and shared hapaxes (words appearing only once in one examinee's text and reappearing only once in another examinee's text) were applied to two known instances of cheating on two different translation tests. It was hypothesised that the collocations and shared hapaxes techniques would be able to determine whether cheating occurred in both instances. Furthermore, the application of successful techniques to known instances of cheating was intended to determine the amount of similarity between two texts necessary to provide significant evidence of cheating and guide the development of an FBI cheating detection policy.

Literature review

The American Heritage® Dictionary defined cheating within a legal context as, a 'fraudulent acquisition of another's property' (nd). The primary concern was whether or not the examinee was the sole author of a translation production, or whether it was plagiarised from another examinee. Authorship attribution techniques examine different linguistic and non-linguistic factors to determine whether two documents are likely composed by the same author.

One of the earliest works on this topic was written by John Allen (1974), who was interested in disputed authorship of literary works and saw the usefulness of incorporating computer science into his analyses. Allen explored the notion of variance within a text and between different texts, hypothesising that texts with greater variance were written by different authors, but those with less variance might originate from a singe source. The analyses used, such as word length, sentence length, part of speech distribution, and word length by part of speech, were the starting grounds for determining authorship.

Finegan (1990) focused on defining an author's style based on linguistic features found in writings that he considered could be uniquely characteristic of the author. Finegan claimed that his examination of the use and misuse of grammatical elements to determine an author's style were replicable, and would meet the requirements for acceptable scientific evidence in United States

Federal Courts, as established under *Daubert v Merrell Dow Pharmaceuticals, Inc.* (1993). Finegan's advances received mixed acceptance by other researchers in the field. Solan and Tiersma (2004) were sceptical of his claims, and referred to Finegan's analyses as the 'eclectic approach', defined as the use of a combination of various features to establish style. Many linguists believed that this approach could be developed into a reliable model, but there was little evidence of proof as yet (Tiersma and Solan 2002). The reliability of stylistic analysis was questioned and the technique rejected in *United States v Clifford* (1983).

In addition to the eclectic approach, Solan and Tiersma discussed the diagnostic and corpus approaches in their 2004 review of developments in Forensic Linguistics. The diagnostic approach counts frequencies of particular features to identify authorship. McMenamin (1993) favoured the diagnostic approach, and postulated a number of quantitative and qualitative methods to determine linguistic patterns unique to a particular author in his 1993 work, *Forensic Stylistics*. Despite adoption by some organisations, McMenamin's techniques met harsh criticism by some linguists. Crystal (1995) noted that McMenamin's work gave the false impression that Forensic Stylistics was an established field. Furthermore, Crystal faulted McMenamin for grounding his analyses in impressionistic statements, lacking arguments in support of his methodology, and drawing vague conclusions, based on evaluative statistics instead of tests of statistical significance.

Chaski (2001) found that many of McMenamin's techniques were unreliable when replicated, with the techniques based on syntactic patterns producing more accurate results than those based on readability, lexical richness, content, or analyses of distinct types of errors, to include, punctuation, morphology, spelling, and grammar. Grant and Baker (2001) claimed that Chaski's analyses were better suited to establish distinctive authorship, determining that two written products had different authors, rather than singular authorship, determining that two written products had the same author.

According to Solan and Tiersma (2004), the greatest hope for authorship attribution rested in corpus linguistics, provided that the corpus of texts was relevant. The use of corpora in authorship attribution gave linguists additional tools for examining language samples. For the first time, linguists accounted for relative word frequencies, collocations, and information on grammatical choices (Coulthard 1994).

Authorship attribution techniques were used to identify cases of plagiarism both within and outside the courtroom context. Johnson (1997) was one of the first to do so by comparing three student texts to three independently produced works in her investigation of plagiarism. She found that common types and tokens between text pairs gave stronger evidence of cheating than type/token ratio or percentage of hapaxes. Olsson's (2004) approach recommended searching for a maximum string of coincidence within a set document, or a percentage of words, and suggested that more than 30% overlap would constitute

plagiarism. Johnson's (1997) and Olsson's (2004) works were relevant to the translation test context because in these situations, plagiarism resulted when the cheater replicated part of another text without giving credit to the author.

One of the few articles on the plagiarism of a translation was Turrell's (2004) work. Turrell reconsidered the evidence presented in an intellectual theft case, where the plaintiff claimed the defendant plagiarised her translation of William Shakespeare's *Julius Caesar* into Spanish. The court reviewed the analyses of overlapping vocabulary, shared once-only words, hapaxes, and shared once-only translation units. The forensic linguist was accepted as an expert witness in the Spanish court and his analyses showed the large amount of overlap, and peculiar overlap, in the texts. Authorship attribution techniques applied to translations aided in the plaintiff winning the case.

Issues

Although authorship attribution techniques used for plagiarism detection may be applied to translation texts, researchers should recognise some of the challenges posed by these techniques. The American Heritage® Dictionary (nd) defines the transitive verb 'plagiarize' as, 'to use and pass off (the ideas or writings of another) as one's own'. This term often refers to writings such as papers or publications, which are meant to express original ideas, and not literary or verbatim translations. A literary translation renders the emotional impact of the source document; whereas, a verbatim translation renders a full and complete meaning in idiomatic target translation without additions or submissions. By definition, the content of a verbatim translation is not original, only its formulation in the target language is original. Therefore, multiple verbatim translations of one source document, as in translation tests, should be very similar to each other. Lexical and phrasal overlap among verbatim translations are likelier than in free-form translations of essays or other literary genres. In addition, good translators attempt to replicate the original style of the author, which may limit the effectiveness of plagiarism detection techniques that focus on the consistency of style within a text.

Furthermore, the analysis process must take into account that the texts examined are translation tests, and not published literary translations, as in Turrell (2004). On the one hand, when different translators render texts that have less idiomatic language, the results tend to be more uniform, making it more difficult to determine whether or not a product was plagiarised. On the other hand, the wide variety of translation skill in the pool of examinees tends to increase variation among the test responses, giving the investigator additional discriminating features to analyse, such as grammatical errors. Incomplete or poor translations may not have significant overlap, and may contain odd usages that are more distinguishable.

With all this in mind, this research considered the following questions:

1. Do the authorship attribution techniques of lexical overlap, type/token ratio, richness score, hapax percentage, collocation, and shared hapaxes effectively support a claim of cheating by plagiarism in FBI Urdu or Arabic translation tests?
2. If so, how similar do two translations of one text need to be in order for them to be considered plagiarised?

Methodology

Data

Two pairs of test responses suspected of cheating were selected for analysis, one pair from Arabic tests and one from Urdu tests. Additionally, a corpus of other Arabic test responses and a corpus of Urdu responses were assembled for comparison. Each corpus was comprised of 40 test productions: the two suspect productions (SPs) and 38 randomly selected productions (RPs). Each corpus of 40 was in turn drawn from the more than 1,500 translation tests the FBI administers annually.

The two tests that were involved in the study are the Arabic Translation Test (TT), which was an assessment of verbatim translation ability from Modern Standard Arabic into English, and the Urdu Verbatim Translation Exam (VTE), which was a similar assessment, but from Urdu to English. The Arabic TT consisted of two passages, producing a translation of approximately 291 total words in 60 minutes. Urdu VTE examinees had 90 minutes to translate four passages, averaging a total of 352 words. Examinees were permitted to use published dictionaries during the translations.

The tests were rated for accuracy, the ability to fully convey the content and meaning of the source passage, and for expression, the ability to write appropriately in the target language, with reference to the Interagency Language Roundtable (ILR) Skill Level Descriptions for Translation Performance (ILR 2007). Accuracy and expression were each assigned independent scores; the lower of the two scores was the final grade.[1]

The raters identified the SPs during the course of normal test scoring, and remarked that they believed the SPs were too similar to have been produced independently.[2] As in all instances of suspected cheating, enquiries were made to find out whether or not there was opportunity to cheat as, for example, whether the examinees took the test at the same time and place. Records indicated that the pair of examinees that produced the SPs did indeed take the test in the same room, sitting in close proximity to each other. In both of these instances, it was determined that the test proctor was occupied with other tasks, and did not consistently monitor the examinees during the entire test period.

Descriptive statistics

As a first examination of the data, descriptive statistics for the Arabic and Urdu corpora were reported in Table 1. The mean word count of all responses given on the Arabic TT was 265.58, with a standard deviation of 41.41 and a standard error of means of 6.55. The median was 277.50, which was quite near the upper bound of the 95% confidence interval for the mean at 278.82. The similarity between the upper bound confidence interval for the mean and the median was an indication that this data may not be normally distributed, leading to a further investigation of skewness and kurtosis by determining the z-score for each. The z-scores for skewness and kurtosis were −3.45 and 2.00 respectively. Since both had an absolute value of greater than 1.96 to ensure normality at $p < .05$, both violate the normality principle, resulting in the transformation of such scores before being used in parametric tests, such as ANOVA.

The mean word count of all responses given on the Urdu VTE was 352.00 words, with a standard deviation of 31.36 and a standard error of means of 4.96, shown also in Table 1. The median was 349.50, quite comfortably in the centre of the 95% confidence interval for the mean, between 341.97 and 362.03. Initial indices for the normal distribution for these data point toward the data being normally distributed. The z-score was determined to be 1.09 for skewness and 0.08 for kurtosis in the Urdu data. Both z-scores had an absolute value less than 1.96, indicating that they were not significant at $p < .05$, and do not indicate a significant deviation from the normal distribution, and do not need to undergo transformation before further analysis with parametric tests.

Additionally, further tests were conducted to examine the characteristics of the data. Kolmogorov-Smirnov tests give further information on the normality of the data. The Arabic data, $D(40) = 0.20$, $p = .000$, were significantly non-normal; whereas, the Urdu data, $D(40) = .09$, $p = .020$, were

Table 1 Descriptive statistics of Arabic and Urdu corpora

Statistic	Arabic	Urdu
Mean	265.58	352.00
SE of Mean	6.55	4.96
Median	277.50	349.50
Variance	1715.00	983.13
SD	41.41	31.36
Minimum	138	290
Maximum	322	427
Range	184	137
Skewness	−1.29	.41
SE of skewness	.37	.37
Kurtosis	1.47	.06
SE of kurtosis	.73	.73

considered to be normal. Homogeneity of variance would have been determined through Levene's test, but required a dependent variable. Due to the lack of a dependent variable in this data, Levene's test was not employed. Instead, the deviations of the two sets of data were examined in relation to their means and ranges. Even though the mean response to the Urdu VTE was approximately 85 words longer than the Arabic TT, the standard deviation was approximately 10 words fewer, indicating that the range in length of response was much broader in the Arabic TT than the Urdu VTE. The different nature of the distributions may be due to the difference in lengths between the two tests or the shorter time permitted to complete the Arabic TT, leading to greater instances of incomplete responses. Nevertheless, the 1.74 variance ratio of the Arabic data to the Urdu data was under 2, indicating it was safe to assume homogeneity of variance.

Authorship attribution techniques

Authorship attribution techniques were chosen for analyses based on the following factors: a) prior use for determining authorship in translations; b) appropriateness of technique for use with translation testing; and c) availability of text processing or statistical analysis software.[3] Techniques selected were:

1. *Lexical overlap* (Allen 1974, Coulthard 2004, Johnson 1997, Olsson 2004, Turrell 2004) In this analysis, each of the 40 test samples for each language were compared to each other, yielding 780 comparisons for Arabic and 780 for Urdu. The number of shared words was reported.
2. *Type/token ratio, hapax percentage, richness score* (Allen 1974, Coulthard 2004, Johnson 1997, Turrell 2004) Type/token ratio was the ratio of type (total number of words produced) and tokens (total number of different lexical items). Hapax percentage was the ratio of the total number of words used only once to total types. The richness score was comprised of a formula that calculated the relationship between types, tokens, and hapaxes (100*log(tokens)/(1-(hapaxes/types))).
3. *Collocations* (Coulthard 2004, Olsson 2004) Collocations examined the number of shared words within strings of different lengths. Strings of four words, six words, eight words, and 10 words were investigated.
4. *Shared hapaxes* (Allen 1974, Turrell 2004) A particular test's hapaxes that occur only once or twice within the entire corpus were identified and analysed to determine if they were shared with another text, and if there were any patterns that occurred between two corpus members.

Although considered within Forensic Linguistics as a separate analysis, lexical overlap, as defined here, was considered a type of collocation. As mentioned, collocations examined the number of shared words within strings of different lengths, and lexical matching software programs allowed the

researcher to select the word length of strings for the program as a parameter to search for matches. If four-word string matches were to be selected, such a computer program would search between two texts for a group of at least four words together that were exactly alike and in the same order. Once all these matches were located, the program then calculated how many words were involved in the matches, with each word only contributing to the count once. Examining lexical overlap was no different from examining collocations of one-word string matches; it simply limited the length of the string it searched to one word, therefore counting the number of words that match exactly. Accordingly, lexical overlap was referred to as a one-word match.

Figure 1 Collocations

1. One-word match: matches exact words
   a. Hamsters *are cute*. (2: 67%)
   b. Hamsters *are* really *cute*. (2: 50%)
2. Four-word match: total words in strings at least four words long
   a. I usually *find her jokes very funny*. (5: 63%)
   b. Most of the time, I *find her jokes very funny*. (5: 50%)

In Figure 1, we can see an example of how collocations were calculated. A one-word match summed the number of words that exactly matched between two texts. In 1, there were two words that matched between sentence a and sentence b: 'are' and 'cute'. The one-word match result for these texts was two. For sentence 1a, two words represent about 67% of the sentence. For sentence 1b, two words represent only 50% of the sentence. When matches between two texts represented a larger percentage of one text, fewer unique words were left.

The sentences in number 2 of Figure 1 modelled a four-word match, the total number of words in strings being *at least* four words long, possibly longer. The collocation or phrase 'find her jokes very funny' was found in both 2a and 2b. Since this phrase was five words long, the four-word match result between 2a and 2b was five. Five words represented about 63% of the total words of sentence 2a, and 50% of the words in 2b.

Choices in WCopyFind

The freeware program WCopyFind was selected to run the collocation analyses in the study. WCopyFind was plagiarism detection software developed at

the University of Virginia so that professors could determine whether or not their students were plagiarising each other's papers or their own papers by turning in the same paper multiple times. Unlike other plagiarism software, it did not connect to the internet to look for matches with sources published there; it only compared texts to other texts either previously or concurrently loaded into the system. The lack of internet connection made this program more suitable for detecting overlap between different tests, because it was not expected that the examinee would have had access to the internet while exposed to the test passages. Before entering texts into WCopyFind, several choices were made to guide the software to select the relevant data when finding matches between texts. The selection categories in the software pertained to punctuation, numbers, letter case, long words and imperfections, and were rationalised according to the needs of the research.

Punctuation

WCopyFind required users to select whether or not the program should ignore outer punctuation, defined as marks placed to the left or the right of a word. Since the study was primarily interested in *lexical* matches within strings, the choice was made to ignore punctuation. Moreover, while entering the texts into the corpus it was noted that examinees often did not follow American conventions for punctuation, resulting in many variants. For the most part, variations in punctuation style did not alter the meaning or effectiveness of the translation. Considering punctuation would have also resulted in increased false negatives, where two passages were exactly alike except for acceptable variation in punctuation.

Numbers

WCopyFind included an option to ignore numbers in texts. Whereas numbers may not have been key elements in student essays, they were considered key elements in translation tests. Rendering numbers from other source languages into English could be often difficult, as other languages use different numerical systems. Therefore, numbers were critical to translations and were retained in all analyses.

Letter case

Both the Urdu and Arabic translation tests were paper and pencil tests, and therefore had to be typed into an electronic document to be processed. Examinees had varied handwriting styles, and sometimes used capital letters. For example, instances occurred where the examinee used the capital letter 'R' even if it were in the middle of a word, as in 'fiRst'. Some examinees wrote using only capital letters for their entire responses. Since, as in the case of punctuation, the use of letter case rarely obscured the meaning of a translation, ignoring letter case was considered justified.

Long words

WCopyFind included the option to ignore recent conventions that were classified as long words, such as email or internet addresses. Because no such items occurred in the Arabic and Urdu translation tests, the option to include long words was not considered important.

Imperfections

WCopyFind included an option where users could allow a certain number of imperfections within the data. Allowing no imperfections would mean that the words within two collocations would have had to match exactly to qualify, referred to as a 'basic' match. Otherwise, the user could choose to allow a certain number of words with imperfections within a perfectly matching phrase, referred to as a 'total' match. In total matches, the program skipped over the allotted number of non-matching words and ignored them when counting words included in strings. For example, the two sentences in Figure 2 would have not qualified for any six-word 'basic' match, because there were not six words in a row that perfectly matched between them. However, a six-word total match run on the two phrases, allowing for two imperfections, would have resulted in the italicised words as matches, ignoring the misspelling of 'trying' and the abbreviation of 'graduate' to 'grad'. 'Finals' to 'for' accounted for a six-word string with one imperfection, as does 'are' to 'Linguistics'. The string of words from 'a' to 'students' was a six-word string with two imperfections. In total, there were eight matched words from all possible six-word strings.

Figure 2 Allowing for errors

1. *Finals are always a* trying *time for Linguistics* grad *students*.
   a. (Six-word total match, 8: 80%)
2. *Finals are always a* triing *time for Linguistics* graduate *students*.
   a. (Six-word total match, 8: 80%)

The reasoning behind allowing for imperfections was to recognise that there were often minor editing (or copying) errors between two similar texts. Allowing two imperfections per string, recommended by the program, was considered reasonable for cases of cheating, as one examinee might not have been able to copy clearly and exactly from another's paper.

The choice of using the 'total' match, over the 'basic' match was further justified by a repeated measures ANOVA, in which both 'basic' and 'total' match results were compared across four-, six-, eight-, and 10-word collocations. One-word matches were not considered because imperfections were

only allowed within strings of words. The results for a one-word total match and a one-word basic match were exactly the same. Data used in this analysis do not meet the qualifications for parametric tests because as the length of the collocation increases, the data become more and more skewed and leptokurtic. The analyses of longer collocations grouped the RPs more closely together and increasingly separated the RPs from the SPs, explained by the fact that there were many fewer matches among the RPs, but not between the SPs. Because data were not parametric, results should be interpreted carefully.

The first step in the one-way repeated measures ANOVA was to determine sphericity with Mauchly's Test, as even small departures from sphericity can produce a large variation in the F value. Mauchly's Test returned a result of $\chi^2(5) = 397.72$, $p = .000$ for the Arabic data, indicating a significant result and a violation of the assumption of sphericity. Similarly, the Urdu data violated sphericity, $\chi^2(5) = 270.54$, $p = .000$. Because the assumption of sphericity was violated, the F value was corrected for both sets of data.

For both Arabic and Urdu, the Huynh-Feldt correction value was not significant and was at least as large, if not larger ($\varepsilon = .81$ for Arabic, $\varepsilon = .75$ for Urdu) than either Greenhouse-Geisser ($\varepsilon = .81$ for Arabic, $\varepsilon = .74$ for Urdu) or Lower-bound corrections ($\varepsilon = .33$ for Arabic, $\varepsilon = .33$ for Urdu), justifying its use. The Huynh-Feldt correction for Arabic results was $F(1) = 786.23$, $p < .05$, indicating that there was a significant difference in whether or not 'total' or 'basic' match was used in the collocations analysis. Likewise, the Huynh-Feldt correction for Urdu resulted as $F(1) = 1112$, $p < .05$, indicating that, here too, there was a significant difference in whether or not 'total' or 'basic' match was used in the collocations analysis. In conclusion, the use of total match was chosen, since allowing for minor differences in longer strings significantly and systematically reported more matches and resulted in a larger quantity of data in the analysis, making it more robust. In the end, two imperfections were permitted.

A priori determinations

There was little precedent for the application of Forensic Linguistic techniques to the field of Language Testing; therefore, limited guidance was offered in the literature for decisions made prior to conducting analyses as well as the interpretation of results. In a sense, present research functioned as a validation study for the application of said techniques. Consequentially, *a priori* decisions on how to manage results were largely informed by a combination of standards for statistical measures and previous research. It was foreseen that *a priori* determinations, such as how an outlier was defined and how to measure proportions, would have required adjustment as research informed practice.

Collocation technique outliers

In the collocations technique, a claim of cheating was supported by determining whether or not any of the Forensic Linguistic techniques examined in this study marked the two SPs as outliers from the rest of the corpus. This theory was supported by the fact that within any random selection of translations of a single source text, there would be factors unique to that production, and a certain amount of overlap between any two productions. If one examinee were to have copied from another, or if both were to have shared answers, a comparison of two such texts would have revealed that they were more alike than any two other naturally occurring productions in the corpus. The analysis would then determine whether the SPs were outliers from the rest of the corpus in shared words or phrases.

Generally, a z-score statistic of greater than 3 or less than −3 would be considered to be an outlier. In the case of the SPs, z-scores for different analyses of greater than 3 were expected due to the nature of the verbatim translation, revealing a good deal of similarity between texts (rather than considerable difference between texts, indicated by a z-score of less than −3). The *a priori* definition of an outlier was a z-score of greater than 3. Complicating the issue was the possibility that text features, specifically a translation passage's difficulty level, may have had the tendency to have caused test responses to be more or less homogeneous. The more homogeneous the translations of a certain passage tended to be, the more difficult it would have been to distinguish any outliers.

Average text length vs individual text length

As previously indicated in Table 1, the test responses of the examinees varied quite a bit in length, ranging from 138 to 322 words for the Arabic TT and from 290 to 427 words in the Urdu VTE. Initially, z-scores were determined using the average length of all text productions in the corpus, instead of the different lengths of individual texts. After initial analyses were run, a concern that results were being 'washed out' by averaging was raised. A series of t-tests were conducted to determine whether or not the proportion of the test comprised by the raw score should have included the average of all texts or the average text length of the two texts being compared.

Regardless of collocation length, the value of t was significant, demonstrating the importance of using the average text length of the two compared texts in determining the z-score rather than the average lengths of all texts. For a one-word match, proportions calculated with only the two texts produced significantly different results than those calculated with the average of all texts ($t(779) = -112.53$, $p < .05$, $r = .97$). Similarly significant results were found for all lengths of collocations sampled from the Arabic data: four-word match: $t(779) = -55.15$, $p < .05$, $r = .89$; six-word match: $t(779) = -32.75$, $p < .05$, $r = .76$; eight-word match: $t(779) = -24.60$, $p < .05$, $r = .68$; and 10-word match: $t(779) = -20.10$, $p < .05$, $r = .58$. Effect sizes reported were moderate to large.

In the Urdu data, significant results were only found in the six-word match ($t(779) = 2.05$, $p < .05$, $r = .07$) and the eight-word match ($t(779) = 3.78$, $p < .05$, $r = .13$). Though results were significant, $p = .04$ for the six-word match, which approached non-significance, and the effect size of .07 was small. Likewise, the eight-word match effect size of .13 was considered small. In the one-word match, proportions calculated with the average length of all texts involved instead of the sample mean did not produce significantly different results than those calculated with the average of all texts ($t(779) = .77$, $p > .05$). Non-significant results recurred in the four-word match ($t(779) = .68$, $p > .05$) and the 10-word match ($t(779) = .76$, $p > .05$). Consequentially, each individual's production length was used in determining its own z-score for both Arabic and Urdu.

Turrell (2004) reported that the six-word match was the determining collocation length for cheating when analysing translations, which was significant for both Arabic and Urdu in the current study. It was curious that the t statistic was consistently significant across all lengths of collocations with high effect sizes for Arabic, yet was only significant for six-word and eight-word collocations for Urdu. One explanation for this phenomenon was that the range of response lengths produced by the Arabic examinees (184) was larger than that of the Urdu examinees (137), even though the mean response length for the Urdu examinees ($M = 352.00$) was greater than the Arabic examinees ($M = 265.58$). Because there was greater variance in the Arabic test scores, it was more important for z-scores to have been calculated from the individual productions involved in the comparison instead of the mean length of all productions in the corpus. Collocations considered proportions calculated as a percentage of the production length of the texts involved in each comparison of productions.

Results

In order for cheating to have been determined, the SPs needed to have presented characteristics that distinguished them from the typical characteristics of other RPs in the corpus. The SPs would have shown much greater similarity and less originality than the RPs, which were all independently produced. Authorship attribution techniques that indicated that only the SPs were outliers would have been considered successful, and only if there was a significant difference in the results for both the SPs than for the rest of the corpus.

Type/token ratio, hapax percentage, richness score

With the exception of lexical overlap, the techniques of type/token ratio, hapax percentage, and richness score were among the most traditional methods used in authorship attribution. Forensic Linguists such as Allen (1974) hypothesised that 'original' or independently produced texts would

have had more total terms, variation within lexical choice, and more once-used words.

Table 2 Type/token ratio, lexical richness score, hapax percentage for Arabic

Applicant	Tokens	Types	Type/token ratio	Richness score	Hapaxes	Hapax percentage
SP 1	295	156	1.89	987.93	117	39.66%
SP 2	291	156	1.87	1011.49	118	40.55%
RP 1	292	166	1.76	998.18	125	42.81%
RP 2	264	142	1.86	838.70	101	38.26%
RP 3	265	140	1.89	942.37	104	39.25%
RP 4	289	147	1.97	882.32	106	36.68%
RP 5	289	165	1.75	966.78	123	42.56%
RP 6	287	149	1.93	832.33	105	36.59%
RP 7	318	163	1.95	849.78	115	36.16%
RP 8	278	153	1.82	984.05	115	41.37%

Allen (1974) hypothesised that authors developed and maintained an individual sense of writing style. He claimed that one way to measure such a style was the ratio between types, the number of different words used, and tokens, the total word count. For these analyses, the SPs and eight RPs were compared. Types and tokens were counted, and the ratio between them determined. In the Arabic data, the SPs were not distinguishable from the other examinees by the measure. Neither of the SPs included the most types or tokens in the sample. Furthermore, they did not have the same type/token ratio, nor did they have type/token ratios that distinguished them from the rest of the corpus. The 'originality' the SPs demonstrated by these techniques was more or less average compared to the RPs.

Allen's (1974) hapax percentage analyses examined the proportion of total words in a production that were only used once, the hapaxes. The SPs produced about the same number of hapaxes as other examinees. SPs 1 and 2 had a hapax percentage of 39.66% and 40.55% respectively, which fell comfortably within the range of hapax percentages from the other examinees, 36.16% to 42.18%. Again, the SPs were not distinguishable from other sample members by examination of hapax percentages.

The third measure that Allen (1974) proposed was the richness score. As in the previous techniques, the richness score, calculated as 100*log(tokens)/(1-(hapaxes/types)), did not distinguish the SPs from the rest of the corpus or give evidence that the SPs were more related to each other than any of the other members of the sample. SP 2 had the highest richness score of the sample, but the richness score for SP 1 was less than that of RP 1. In the Arabic data, no determinations of authorship can be concluded from the richness score.

As in the Arabic data, none of Allen's (1974) analyses distinguished either

Table 3 Type/token ratio, lexical richness score, hapax percentage for Urdu

Applicant	Tokens	Types	Type/token ratio	Richness score	Hapaxes	Hapax percentage
SP 1	324	197	1.64	915.88	143	44.14%
SP 2	365	211	1.73	932.14	153	41.92%
RP 1	382	222	1.72	1023.60	166	43.46%
RP 2	324	204	1.59	1249.15	163	50.31%
RP 3	326	199	1.64	1064.11	152	46.63%
RP 4	337	194	1.74	891.56	139	41.25%
RP 5	339	195	1.74	865.59	138	40.71%
RP 6	342	205	1.67	998.99	153	44.74%
RP 7	383	224	1.71	1180.89	175	45.69%
RP 8	306	189	1.62	999.58	142	46.41%

of the Urdu SPs from the rest of the corpus or resulted in similar scores for the two SPs. SP 1 and 2's type/token ratios of 1.64 and 1.73 respectively were not unlike the ratios for the other independently produced texts ranging from 1.59 to 1.74, as shown in Table 3. Likewise, their hapax percentages, 44.14% and 41.92%, were not extreme scores outside the sample's range of results, from 40.71% to 50.31%. The SPs' richness scores, at 915.88 and 932.14, were neither the highest nor the lowest of the range produced by the sample, which was from 865.59 to 1249.15. Again it was seen that these techniques offered no evidence to distinguish the performances of the SPs from the performances of the other samples in the data sets.

Lexical overlap and collocations

Arabic and Urdu test data

The lexical overlap and collocations analyses examined the number of shared words within strings of varying lengths. Initially, the entire Arabic TT production was taken as a whole passage, combining the different passages within the test to create one document in an attempt to look at shared words across the entire text. The same process was done for the Urdu VTE. WCopyFind compared each of the 40 productions in the corpus with each of the other tests, resulting in 780 total comparisons. This analysis was run for strings of increasing length: one-word strings (lexical overlap), four-word strings, six-word strings, eight-word strings, and 10-word strings. The 780 comparisons were ranked in ascending order by the z-score, and the top five comparisons for each collocation length were reported in Table 4.

The column 'Total words matched' reported the number of words included within the collocations of different lengths. 'Percentage of total texts matched' referred to the percentage of average text length of the two texts in the comparison that was included in the 'Total words matched'. The resulting

percentages were converted to z-scores in 'Z-score', and were ranked out of the 780 different combinations of examinees in 'Rank'. The comparison that included the two SPs was marked in bold, italicised print.

Table 4 Collocation technique results

Match string length	Arabic				Urdu			
	Total words matched	Rank	Percentage of total texts matched	Z-score	Total words matched	Rank	Percentage of total texts matched	Z-score
1-word string	***247***	***1***	***84%***	***3.76***	***251***	***1***	***73%***	***3.29***
	209	2	72%	2.20	257	2	71%	2.92
	192	3	71%	2.04	244	3	69%	2.44
	202	4	70%	1.89	235	4	68%	2.37
	206	5	70%	1.88	240	5	68%	2.29
4-word string	***237***	***1***	***81%***	***6.47***	***184***	***1***	***53%***	***6.32***
	137	2	47%	2.98	143	2	40%	3.99
	128	3	46%	2.87	134	3	35%	3.23
	126	4	45%	2.8	125	4	35%	3.18
	126	5	45%	2.76	125	5	33%	2.87
6-word string	***232***	***1***	***79%***	***9.02***	***148***	***1***	***43%***	***6.88***
	114	2	41%	4.09	92	2	26%	3.57
	103	3	37%	3.58	93	3	26%	3.44
	101	4	36%	3.48	91	4	26%	3.42
	94	5	34%	3.12	84	5	24%	3.18
8-word string	***219***	***1***	***75%***	***10.10***	***116***	***1***	***34%***	***6.76***
	96	2	35%	4.19	85	2	24%	4.54
	90	3	32%	3.88	80	3	22%	4.02
	87	4	31%	3.72	76	4	21%	3.78
	82	5	30%	3.55	68	5	21%	3.66
10-word string	***219***	***1***	***75%***	***11.47***	***98***	***1***	***28%***	***6.70***
	96	2	35%	4.92	71	2	20%	4.34
	90	3	32%	4.57	62	3	19%	4.11
	78	4	28%	3.86	59	4	17%	3.61
	73	5	26%	3.53	60	5	17%	3.55

Results indicated that even in the one-word match, the two Arabic SPs had the highest number of total matched words, and the matched words accounted for an average of 84% of their total productions, with a corresponding z-score of 3.76. This z-score was already greater than the *a priori* prediction of 3 for determining outliers; the comparison of the SPs was the only one that qualified as an outlier under the definition. The next most similar pair of productions had a z-score of 2.20. Among four-word collocations, 81% of the SPs matched. This statistic had a corresponding z-score of 6.47, which was almost twice the z-score for the one-word match. The next closest z-score was 2.98, which fell just short of outlier status. In the six-word strings, the percentage of word matches between the SPs was 79%, but the

z-score jumped to 9.02. The next closest z-score was less than half as much, at 4.09, but would have now, along with several other combinations behind it, qualified as an outlier with a z-score greater than 3. The same trend was repeated in the eight-word match and the 10-word match. The SPs showed a combined percent of 75% in both cases, and their z-scores were 10.10 and 11.47 respectively. The z-scores of the next closest combinations also crept up as the length of the collocations increased, but never above 5. The Arabic test SPs maintained a z-score of over double any other comparison of texts.

Table 4 also showed that the collocations analyses for the Urdu test SPs resulted in a 73% sharing of the total production in the one-word match. The corresponding z-score of 3.29 gave the SPs outlier status, according to the traditional definition. Of note was the fact that the next closest comparison of productions had a z-score of 2.92, which was not very different from SPs' z-score, and almost in outlier status. For four-word collocations, the SPs again received a high combined percentage of their total productions that matched, at 53%. This statistic had a corresponding z-score of 6.32, much higher than the z-score for the next closest pair of productions, at 3.99. Pairs ranked 2 through 4 were all higher than 3, and qualified as outliers as well. In the six-word match, the combined percentage dropped to 43%, with the z-score remaining in the high 6 range, at 6.88. The next closest z-score was just over half as much, at 3.57, and it still, along with several other pairs behind it, qualified as an outlier with a z-score greater than 3. Results from the eight-word and the 10-word matches did not change much for the SPs, with z-scores of 6.76 and 6.70 respectively. As in the Arabic data, the z-scores of the next closest combinations also crept up as the length of the collocations increased, but again, the z-scores of pairs of the RPs never reached 5.

Subsequently, the question was raised as to whether or not the traditional definition for an outlier, the absolute value of the z-score greater than 3, should be amended for collocations analyses. Adjusting the definition of outlier to the absolute value of a z-score greater than 6 would have qualified only the comparison of the SPs as outlier, and not any comparisons of RPs. Recall that the SP examinees had the opportunity to cheat. Further investigation into the proctoring procedures of other examinees determined that none of the pairs of examinees whose comparisons ranked in the top five for any string length for Arabic or Urdu took the test on the same day, so there was no circumstantial evidence for cheating for any of the RPs. Therefore, comparisons of independently produced texts may have resulted in z-scores near 5, but not 6, justifying a z-score of 6 to determine cheating.

Four examinees within the same day

As mentioned previously, the SP examinees from both the Arabic and Urdu data had the opportunity to copy from each other. The only difference in the test administration of the SP examinees and the RP examinees who took the

test together was that the proctor was not consistently attentive to the SP examinees. Nevertheless, the assumption that examinees who tested in the same sessions would submit original productions, i.e. not cheat, was investigated. Accordingly, four Urdu RPs from the corpus who took the test at the same time were examined for similarities. Results of all possible pair comparisons between these four examinees (labelled examinees a through d) were reported in Table 5. The results from the comparison of the two SPs were also indicated.

Table 5 Four Urdu examinees tested on the same day

Paired comparison		1-word %	1-word z-score	4-word %	4-word z-score	6-word %	6-word z-score	8-word %	8-word z-score
Examinee a	Examinee b	65%	1.66	19%	0.49	9%	0.21	4%	−0.36
Examinee c	Examinee b	60%	0.75	16%	0.02	5%	−0.75	3%	−0.53
Examinee c	Examinee a	60%	0.66	24%	1.32	11%	0.43	2%	−0.64
Examinee d	Examinee b	63%	1.29	19%	0.49	11%	0.46	5%	0.08
Examinee d	Examinee a	58%	0.26	24%	1.34	15%	1.39	12%	1.60
Examinee d	Examinee c	59%	0.49	27%	1.85	18%	1.90	14%	2.09
SP 1	SP 2	73%	3.29	53%	6.32	43%	6.88	34%	6.76

Each of these four examinees took the Urdu VTE in the same session on 17 September 2006. Examinees were labelled a through d, and each of the six comparisons was reported. There were no apparent instances of outlier status, using either the traditional minimum of 3 or the amended minimum of 6, for any of the six comparisons of same-day examinees' productions. The highest z-score reported was 2.09 in the eight-word string match comparison of examinees c and d, which did not qualify for outlier status. The two SPs, however, had z-scores greater than 3 regardless of length of the string, and qualified as outliers with a z-score of greater than 6 for the four-word string matches as well as the longer collocations. Based on the data above, having a common setting during translation test administration did not by itself cause test responses to be any more similar.

Urdu test passages

Unlike the Arabic TT, the Urdu VTE was designed so that each of the four passages included within the test would be increasingly difficult to translate. Further consideration as to how this test's characteristics could have affected the collocations analyses led to additional examination of the Urdu test data. Considering that the first passage should have been the easiest to translate, perhaps the translations rendered by the examinees would have been more accurate and with more consistent expression, and therefore more similar to each other than those of the more difficult passages. This increased homogeneity among productions may have made the collocations technique less effective in determining cheating. Consequentially, although cheating might

have occurred, the SPs would have to have been different enough from the RPs to make a clear determination. Likewise, passages that were difficult to translate might have had the opposite effect, making cheating easier to detect due to the greater variance in responses. In order to investigate these possibilities, a passage by passage analysis of the Urdu test was conducted. In Table 6, the top three pairs of examinees that had the most similar responses in the four passages according to the collocation tests of different lengths were reported, in addition to the SP comparison. In cases where the SP comparison, indicated by bold and italicised fonts, were not included in the top three rankings, they were also listed with their applicable rank.

Table 6 Urdu collocations by passage

	Passage 1			**Passage 2**			**Passage 3**			**Passage 4**		
	Percent of total text matched	**Z-score**	**Rank**	**Percent of total text matched**	**Z-score**	**Rank**	**Percent of total text matched**	**Z-score**	**Rank**	**Percent of total text matched**	**Z-score**	**Rank**
1-word	77%	2.47	1	***72%***	***3.27***	***1***	***71%***	***2.13***	***1***	***66%***	***2.64***	***1***
string	77%	2.45	2	70%	3.01	2	70%	2.05	2	65%	2.51	2
	77%	2.43	3	68%	2.77	3	69%	2.04	3	65%	2.45	3
	76%	***2.35***	***5***									
4-word	67%	3.33	1	54%	5.21	1	***63%***	***6.75***	***1***	***46%***	***4.44***	***1***
string	65%	3.21	2	***53%***	***5.08***	***2***	51%	5.31	2	40%	3.61	2
	65%	3.18	3	43%	3.81	3	42%	4.11	3	37%	3.20	3
	58%	***2.57***	***7***									
6-word	59%	3.31	1	44%	6.68	1	***49%***	***7.81***	***1***	40%	4.94	1
string	57%	3.15	2	***43%***	***6.54***	***2***	32%	4.96	2	***39%***	***4.80***	***2***
	57%	3.11	3	32%	4.77	3	29%	4.51	3	30%	3.56	3
	42%	***1.87***	***28***									
8-word	60%	4.13	1	44%	10.51	1	***53%***	***11.61***	***1***	40%	5.94	1
string	59%	4.08	2	28%	6.55	2	28%	5.95	2	30%	4.42	2
	51%	3.32	3	25%	5.85	3	21%	4.35	3	26%	3.75	3
	42%	***2.56***	***8***	***21%***	***4.85***	***4***				***24%***	***3.32***	***12***

As predicted, the collocation analyses of Passage 1 did not rank the SPs in the top three comparisons for any length of collocation examined. The highest rank that the SPs reached was 5 in the one-word match, with a z-score of 2.35, and the lowest was 28 in the six-word match, with a z-score of 1.87. In no collocations did they qualify as outliers, as all z-scores were below 3. In Passage 2, the SPs appeared much higher in the rankings of the top three, but not consistently reaching number 1, an essential criterion to determine cheating had occurred instead of the similarity of the passages occurring naturally. In fact, the SPs were only ranked number 1 in the one-word collocation, with a z-score of 3.27, 72% of the total production matched. Even in the one-word collocation, they were near the number 2 rank at a z-score of 3.01. In the eight-word match, the SPs fell to fourth place, with a z-score of 4.85,

far behind the first-ranked pair, with a z-score of 10.51.[4] Again, in the second passage, there was no clear support for the claim that the SPs cheated.

In the third passage, a different story emerged. Here, the SPs were ranked first for each collocation analysis, regardless of length of string. With a z-score for the SPs of 2.13, the one-word match did not represent them as outliers, but the SPs' z-scores progressively increased as the length of the collocation increased. In the four-word match, the SPs had a combined percentage of 63% and a z-score of 6.75, qualifying them as outliers by both the traditional standard of 3 and the revised standard of 6. Nevertheless, the pair in rank 2 had a very high z-score, 5.31, as did the pair in rank 3. In the six-word match, the z-score for the SP comparison jumped to 7.81. The next closest z-score was less than 6, at 4.96. The eight-word match produced a remarkable z-score for the examinees of 11.61. Once again, it was more than twice the z-score of the second rank.

In the fourth passage, the SPs held the highest rank for similar productions in both the one-word match and the four-word match with z-scores of 2.64 and 4.44 respectively. Nonetheless, another pair surpasses the SPs in amount of similarity in the six-word string match with the top z-score of 4.94, only slightly higher than the SPs' z-score of 4.80. In the eight-word collocation, the SPs drop to 12th place in the rankings, with a z-score of 3.32, much lower than the first-ranked pair's z-score of 5.94. All z-scores reported in the Passage 4 analyses had z-scores below 6, providing further support for the adjusted z-score determination for outlier at 6 for the collocations techniques. Passage 4 offered no evidence for cheating, perhaps because no cheating occurred in this passage, or the copied text was not detected by this technique.

Pearson correlations

In an effort to determine which length of collocation produced the most reliable results in determining cheating, Pearson correlation matrices were compiled for the Arabic and Urdu test data. Percentage of overlap in passages was correlated between collocations of different lengths. Results for both languages were reported in Table 7.

Pearson correlations for Arabic ranged from $r = .50$, for the correlation between the four-word match and the 10-word match, to $r = .92$, for the correlation between the six-word match and the eight-word match. All were considered to be moderately to highly correlated. Considering that the six-word match and the eight-word match were the most strongly correlated, at $r = .92$, there would have been no great gain in using the longer eight-word match, and the six-word match was determined to be sufficient. Pearson correlations for Urdu ranged from $r = .58$, for the correlation between the four-word match and the 10-word match, to $r = .94$, for the correlation between the 10-word match and the eight-word match. Considering that that eight-word match and the 10-word match were most strongly correlated at

Table 7 Pearson correlations of Arabic and Urdu

Collocation length	1-word strings	4-word strings	6-word strings	8-word strings	10-word strings
			Arabic		
1-word strings	–	.87	.72	.66	.56
4-word strings		–	.90	.82	.50
6-word strings			–	.92	.75
8-word strings				–	.86
10-word strings					–
			Urdu		
1-word strings	–	.73	.79	.64	.58
4-word strings		–	.67	.57	.76
6-word strings			–	.85	.86
8-word strings				–	.93
10-word strings					–

r = .94, there would be no great gain in using the longer 10-word match, and the eight-word match would be sufficient. Like the Pearson correlations for Arabic, there was a strong correlation between the six-word match and the eight-word match, at r = .85. With a correlation over r = .80, the six-word match was also considered a sufficiently effective tool.

Shared hapaxes

Although the number and percentage of hapaxes did not give much information about plagiarism, further analysis of unique and shared hapaxes provided more discriminating results. Unique hapaxes were defined as terms that not only occurred only once within a text, but they only occurred once within the corpus. Unique hapaxes existed in every production examined. Many times unique hapaxes were labelled as a lexical choice error in the grammatical analyses, meaning that the word used was very awkward in that context or did not make sense at all. (This sometimes happens when examinees use dictionaries, but choose the wrong option for the context.)

On some occasions, hapaxes were almost unique; they were shared by only one other examinee in the corpus. Typically, this phenomenon of once-shared hapaxes was randomly distributed throughout the pool of other texts, meaning one examinee would not have consistently shared his or her once-shared hapaxes with one other particular examinee. If the sharing of these once-shared hapaxes occurred consistently between the same pair of examinees, then it could be determined that they cheated from each other.

In the case of the SPs, however, they included very few truly unique, unshared hapaxes, and their once-shared hapaxes were almost exclusively

Table 8 Unique and once-shared hapaxes

	Mean	SP 1	SP 2
		Arabic	
Unique Hapax	6.75	2	3
Once-Shared Hapax	3.95	14	16
		Urdu	
Unique Hapax	9.50	0	0
Once-Shared Hapax	4.30	10	10

shared with each other, as seen in Table 8. Approximately 75% of the SPs' once-shared hapaxes occurred in each other's texts. Where the SPs shared 12 of their once-shared hapaxes in the Arabic data, there was no other production that shared more than two once-shared hapaxes with any other RP. Likewise in the Urdu data, the SP shared seven of their 10 once-shared hapaxes with each other; whereas, no other examinee shared a once-shared hapax with the same other examinee more than two times. Consequently, the SPs' sharing of unique hapaxes separated them from the rest of the group, indicating cheating occurred.

To further support evidence given from the shared hapaxes, both the Arabic and Urdu SPs had very few unique hapaxes. SP 1 had only two unique hapaxes and SP 2 had only three unique hapaxes, many fewer than the mean of the Arabic test corpus of 6.75. The Urdu test SPs had no unique hapaxes at all, unlike the mean of the Urdu test corpus, which was 9.50. In both cases, there was support that cheating occurred among the SPs. The SPs had both an abnormally high number of shared hapaxes, and an abnormally low number of unique hapaxes, perhaps indicating that they were sharing their hapaxes, and likely other words as well.

Conclusions

Data analysed in this study indicated that there was an answer to the first research question, 'Do the authorship attribution techniques of lexical overlap, type/token ratio, richness score, hapax percentage, collocation, and shared hapaxes effectively support a claim of cheating by plagiarism in FBI Urdu or Arabic translation tests?'. The authorship attribution techniques of type/token ratio, richness score, and hapax percentage were unsuccessful in distinguishing the performance of the SPs from the rest of the corpus. Perhaps this result was due to the nature of translations; multiple verbatim translations of a single source text shared the content, though not the expression of the content. The inherent similarity between multiple renderings of

a single text may have restricted the amount of variety, even variety in style, possible between different texts. Furthermore, there was little evidence for using lexical overlap as a determination for cheating. Although the z-scores of the SPs in the one-word string matches often were higher than any other comparison of two examinees, the z-scores would barely qualify them as outliers, just over 3, or would qualify other examinees as outliers that could not have cheated.

Alternatively, the collocations technique appeared to be able to substantiate a claim of cheating, meeting certain qualifications. Two minor imperfections within a collocation were permitted in the collocation analyses, and the z-score was set at an absolute value of greater than 6, instead of the standard 3. These alterations to the analysis proved to consistently mark the SPs as outliers from the rest of the corpora, and should be considered as criteria for collocation analyses from this point forward. Considering the high-stakes nature of the test and the severe consequences for the examinee, this stricter standard for outliers would help prevent false positive indications of cheating.

The use of collocations consistently indicated that the SPs were more similar to each other than to any of the other 779 pairs in the corpus when considering the entire test productions for both Arabic and Urdu. Collocations of increasingly longer lengths increased the SPs' z-scores making them stronger outliers, but in the case of Urdu, collocations beyond a six-word match showed decreasing returns. A passage by passage analysis of Urdu revealed that the same results were not consistent across passages. Though this inconsistency certainly did not rule out the use of collocations, it revealed that the SPs may not have cheated on all parts of the test as well as some insights into the nature of translation difficulty. It was unclear whether or not cheating occurred in other passages, or whether the collocations technique was not useful for supporting a claim of cheating in the easier passages. Furthermore, there appeared to be some limitations to the use of collocations on short texts, as most of the individual passages had responses from 75 to 125 words. Whereas length and difficulty of the passages may have been mitigating factors, simply taking the test in the same session with other examinees was not. An examination of RPs that took the test at the same time and in the same location revealed that the similarity between the SPs was probably due to shared information rather than shared environment.

In regards to the second research question, 'How similar do two translations of one text need to be in order for them to be considered plagiarised?', conclusive evidence was established due to the limitations inherent in the data set, namely that only two pairs of SPs were identified. Results from this study suggested the use of collocations of six-word strings (Coulthard 2004, Turrell 2004) produced strong enough evidence of cheating. Strings of longer lengths did not greatly improve the SPs' distinction from the rest of the corpus. Additionally, Pearson correlations revealed that there was high reliability,

$r > .80$, between the six-word, eight-word, and 10-word matches. Therefore, the use of six-word collocations was deemed sufficient. Furthermore, setting the z-score at 6 instead of 3 consistently isolated the two SPs as outliers when used in combination with six-word string matches. Although the results of this study pointed towards the utility of six-word collocations and an elevated z-score for determining cheating, additional evidence from translation tests in other languages would strengthen this claim.

Finally, the once-shared hapaxes technique identified the SPs as outliers, revealing characteristics of those productions that did not occur naturally in the other corpus members' productions. The application of this authorship attribution technique was only applied to the SPs and their corpora, a limited sample, and was thus far only analysed descriptively. Nevertheless, the shared hapaxes technique demonstrated that the SPs resembled each other much more, at least three times more, than any other pair. Further evidence for the use of this technique would show promise in determining cheating in translation testing.

Limitations and further research

One limitation was that the collocation technique did not clearly distinguish whether there was collusion between the examinees or if one examinee copied from another examinee, who was unaware of the cheating that occurred. Further research would aid in making a more reliable determination of which party or parties may have been involved.

Furthermore, computer programs have difficulty considering the nuances of language in performing such analyses. No doubt, programs provide consistent results, but there may have been times when certain words or phrases would need human judgment to be correctly interpreted. It was also difficult to consider where to draw the line when differentiating lexical matches. Would it be right to consider 'goodwill' different from 'good will'? Should the order of clauses within a sentence matter when both are grammatically correct? In order for Forensic Linguistic techniques such as collocations, shared hapaxes and others to be accepted in court, peer reviews and the establishment of error rates for techniques are required under the United States Federal Rules of Evidence. Further research on the application of authorship attribution techniques to language testing would not only benefit test takers and organisations that use test scores, but it also holds the potential to expand our understanding of language.

Notes

1. Accuracy scores are not based on discrete error counts, but on the degree to which a translation fully conveys the meaning of the source text. Accuracy

errors reflect instances where the meaning conveyed is different from that of the source, according to the ILR Skill Level Descriptions for Translation Performance (ILR 2007). Although most final scores result from a lower accuracy score than expression score, a weakness in expression can distort or obscure the meaning conveyed in the translation, therefore affecting the translation's accuracy.

2. Under the current FBI procedures, when there is evidence for cheating, the applicant file is closed, the applicant can never apply to the FBI again, and the instance of cheating is recorded on the applicant's permanent record. Raters currently determine cheating by expert judgment; they score a large quantity of tests and are familiar with typical variations.
3. WCopyFind, plagiarism software developed at the University of Virginia, was used in some analyses. Available at http://plagiarism.phys.virginia.edu/Wsoftware.html
4. Pair ID 2738 is the comparison between two very high scoring examinees' productions. No further analysis is offered here for this outlier at this point, but it is worth investigating in further research.

References

Allen, J R (1974) Methods of author identification through stylistic analysis, *The French Review* 47 (5), 904–916.

Chaski, C E (2001) Empirical evaluations of language-based author identification techniques, *International Journal of Speech, Language and the Law: Forensic Linguistics* 8 (1), 1–65.

cheating (nd) The American Heritage® Dictionary of the English Language, Fourth Edition, retrieved 10 May 2007, from Dictionary.com website: <http://dictionary.reference.com/browse/cheating>.

Coulthard, M (1994) On the use of corpora in the analysis of forensic texts, *International Journal of Speech, Language and the Law: Forensic Linguistics* 1 (1), 27–43.

Coulthard, M (2004) Author identification, idiolect, and linguistic uniqueness, *Applied Linguistics* 25 (4), 431–447.

Crystal, D (1995) Review of the book *Forensic Stylistics* by G R McMenamin, *Language* 71 (2), 381–385.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

Educational Testing Service (ETS) (1999) *Cheating Fact Sheet – RESEARCH CENTER – Cheating is a Personal Foul,* retrieved 10 May 2007 from <http://www.glass-castle.com/clients/www-nocheating-org/adcouncil/research/cheatingfactsheet.html>.

Finegan, E (1990) Variation in Linguists' analyses of author identification, *American Speech* 65 (4), 334–340.

Grant, T and Baker, K (2001) Identifying reliable, valid markers of authorship: a response to Chaski, *International Journal of Speech, Language and the Law* 8 (1), 66–79.

Interagency Language Roundtable (ILR) (2007) Interagency Language Roundtable Skill Level Descriptions for Speaking Translation Performance, *Interagency Language Roundtable*: Washington, DC, retrieved 4 August 2008 from <http://www.govtilr.org/Skills/ILRscale2.htm>.

Johnson, A (1997) Textual kidnapping – a case of plagiarism among three student texts? *International Journal of Speech, Language and the Law: Forensic Linguistics* 4 (2), 210–225.

McMenamin, G R (1993) *Forensic Stylistics,* Amsterdam: Elsevier.
Olsson, J (2004) *Forensic linguistics: An introduction to language, crime and the law,* London & New York: Continuum.
plagiarize (nd) The American Heritage® Dictionary of the English Language, Fourth Edition, retrieved 8 March 2008, from Dictionary.com website: <http://dictionary.reference.com/browse/plagiarize>.
Solan, L M and Tiersma, P M (2004) Author identification in American courts, *Applied Linguistics* 25 (4), 448–465.
Tiersma, P M and Solan L (2002) The linguist on the witness stand: Forensic Linguistics in American courts, *Language* 78 (2), 221–239.
Turrell, M T (2004) Textual kidnapping revisited: the case of plagiarism in literary translation, *International Journal of Speech, Language and the Law: Forensic Linguistics* 11 (1), 1–26.
United States v. Clifford, 074 F.2d 86 (3d Cir. 1983).

6 Cause and effect: the impact of the Skills for Life strategy on language assessment

Philida Schellekens
Independent researcher and consultant

Abstract

In 2001 the government in England launched a strategy to improve the literacy and numeracy skills of the population. This was a brave step in recognition of the poor levels of basic skills identified in studies such as the 1997 International Adult Literacy Survey. The Skills for Life strategy became a major government initiative which has been generously funded and energetically pursued. Government created an infrastructure by introducing the first national standards for literacy and numeracy, national curricula, learning materials and qualifications. National targets were set to monitor the effectiveness of delivery over time. In this paper we shall reflect on how the Skills for Life has fared and what can be learned from the experience, both in terms of expected and unexpected consequences. For example, the achievement of migrants and refugees, who need English for social and work purposes, is assessed against the national literacy standards which were designed for native English speakers. This has had important implications, not all positive, for curriculum design, testing and classroom delivery as well as for the key measure of the government strategy: the collection of data on achievement.

Introduction

In 2001 the government in England launched a strategy to improve the literacy and numeracy skills of the population.[1] This was a brave step in recognition of the poor levels of basic skills found among its native English-speaking citizens. These were first identified in a 1997 study which was based on the OECD (Organisation for Economic Co-operation and Development) International Adult Literacy Survey model and the Moser report two years later. The Skills for Life strategy, as it became known, has since become a major government initiative which has been generously funded. This has allowed for a significant

expansion of learning opportunities as well as the creation of a government-controlled infrastructure: national standards for literacy and numeracy, national curricula, learning materials and qualifications for learners and teachers. Public service agreements have been set to monitor the achievement of targets.

This paper charts the implementation of the Skills for Life strategy, with particular reference to migrants and refugees. This section of the population, new and long-term residents whose first language is not English, was not the primary target group for the strategy. As we shall see, this has had both expected and unexpected consequences for language assessment and the planning and delivery of learning programmes. Before we explore these aspects, first a brief profile of the target group and the terminology that is used to describe English language teaching.

Government-funded provision which is offered to migrants and refugees is called English for Speakers of Other Languages (ESOL). In this paper people whose first language is not English and who attend English language classes are referred to as 'language learners'. Where comparisons are made between categories of learners, e.g. speakers of other languages and people whose first language is English, the former will be referred to as 'second language speakers'. This is for the pragmatic reason that this term is easier to understand than the alternative 'other language speakers'. This does not imply that the learners' other language skills are not important.

A profile of the learners: migrants and refugees

Language learners in England are often categorised into four groups for funding and educational purposes:

- Refugees who have fled national or international conflict. People from Iraq, Sri Lanka and Somalia make up the largest groups, but conflict in Africa, in countries such as Burundi, Zaire, and Zimbabwe, has also caused many to flee.
- New Commonwealth citizens, many of whom came in the 1970s and 1980s for economic reasons from countries such as India, Bangladesh, and Pakistan. Many of the new arrivals in this category come for reasons of family reunion, most often because they have married a partner already living in the UK.
- EU citizens, of whom there has been a major influx since 2004 when countries such as Poland, Slovakia and Lithuania acceded to the European Union. Their entitlement to free ESOL classes was drastically reduced in 2006.
- People from countries outside the EU, e.g. Japan, the Philippines and Argentina. They are normally not entitled to free ESOL provision, unless they have settled permanently in the UK.

It will be clear from the examples given above that there is enormous diversity in background and nationality. There is also increasing diversity in the areas where new arrivals settle. In the past, migration was mainly concentrated in the big metropoles such as Birmingham and in particular London, which Storkey, Maguire and Lewis (1997) chart over time. More recently, migration patterns have become more dispersed. For example, the latest Home Office *Accession Monitoring Report* (2008) records that between 2004 and 2008 Anglia (in the east of England) attracted 120,000 workers from the European accession countries. This makes Anglia the region with the highest percentage of Eastern Europeans for the whole of the United Kingdom.

There are many reasons why learners decide to come to England but they are principally: to seek asylum, to find employment, to join spouses and/or family, to experience living in another country and to make sure that the next generation benefits from educational opportunities. Schellekens (2001) found that, once migrants and refugees have arrived in the UK, their prime reason for learning English is to improve their employment prospects.

Data on migration and language needs

The United Kingdom does not collect data on the number of residents who have English language needs. This has hampered not just the planning for educational provision but also for housing and health facilities. However, two indirect sources give a broad indication of settlement in the UK. The first data set tracks the number of people who have come to the UK since the eight accession countries joined the European Union on 1 May 2004. Since then, the UK has operated an open frontier policy which has resulted in many Eastern European workers entering the UK, making up the vast proportion of recent arrivals. Data collected by the Home Office through its voluntary Worker Registration Scheme show that a total of 812,000 applications were approved between May 2004 and March 2008. Figure 1 gives a breakdown of the nationalities.

It should be noted, though, that the number of applicants does not equal the people who have settled in the UK long-term. This is because people who leave the UK are not required to de-register. Nevertheless, the Home Office data give an idea of the flow of people from Eastern Europe.

The second data set concerns the Office for National Statistics which provides the most recent census data on the number of people born outside the UK (See Table 1).

Clearly, the number of people who were born abroad does not equate to the number of people who have English language needs. Many speak English very well. However, there is a concern that a significant proportion of migrants and refugees lacks the English language skills required to function in society and at work. This is indicated, for example, by the employment

Figure 1 Home Office Accession Monitoring Report May 2004–March 2008

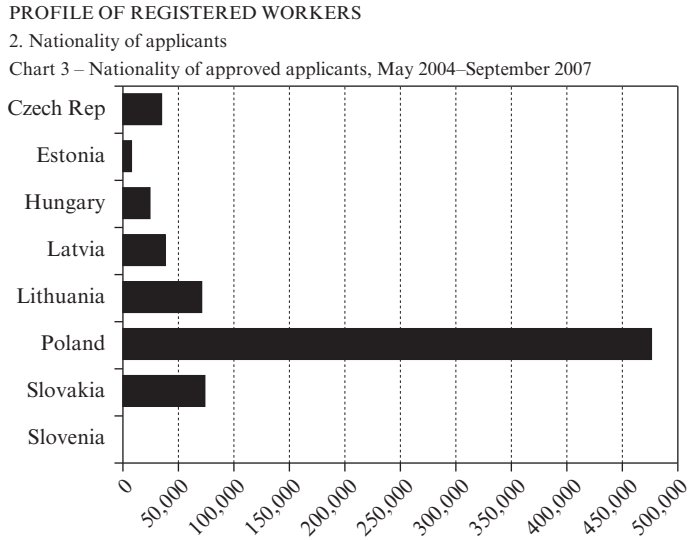


Table 1 Census data on those born in non-English speaking countries

Country	Total population	People born in countries where English is not the national language	%
England & Wales	52,041,916	3,475,507	6.7
Scotland	5,062,011	155,943	3.1
Northern Ireland	1,685,274	18,255	1.1
Total	**58,789,201**	**3,649,705**	**6.2**

Office for National Statistics 2001 Census

rates for ethnic minorities, which are based on the Labour Force Survey. The Ethnic Minority Employment Taskforce report (2008, quarter 2:3) records that employment for ethnic minorities is 60.5%, which is much lower than that for the white population at 76.5%. The 2006 Employment Report states 'Ethnic minorities are still over twice as likely as their white counterparts to be unemployed' (2006:6).

While government reports tend not to provide analysis of the reasons why ethnic minority economic participation is below that of the white population, the inability to use English effectively is surely one of the main factors which holds back second language speakers. Their employment and other skills could be unlocked if they had sufficient proficiency in English. Thus language provision would not just benefit individual second language speakers and

their families but also wider society. The ability to speak English would help integration into their adopted society and finding employment would enable people to contribute taxes rather than have to rely on state benefits.

It will be clear from the census and accession monitoring reports that the inflow of second language speakers has been significant over time. While government departments and education and training providers lack, and sorely need, accurate data on the demand for English language provision, Census and Home Office reports indicate that potentially over four million residents are eligible. Both the size of the target group and the demand for language learning programmes has made the Skills for Life strategy all the more important.

The Skills for Life strategy targets

The Skills for Life strategy was launched in 2001 with the publication of the document *The National Strategy for Improving Adult Literacy and Numeracy Skills*. This document acknowledged the existence of '7 million adults in England who could not read and write at the level we would expect of an 11-year-old' (2001:1). This referred to native English speakers, as second language speakers had not been included in the data collection. The strategy set out to 'deliver radical improvements in standards and achievements' (2001:2):

- national targets for literacy and numeracy of 750,000 learners to achieve literacy and numeracy qualifications by 2004 including '50,000 refugees and speakers of other languages'
- the entitlement to free tuition for 'every adult who is improving his or her literacy or numeracy'
- robust national standards and curricula
- national tests for literacy and numeracy
- new learning materials
- new qualifications for teacher training in literacy and numeracy
- a substantial increase in the budget to fund provision between 2001 and 2004 (2001:10–15).

Overview of achievements 2001–07

The introduction of the Skills for Life strategy has undoubtedly brought many gains. Of real benefit has been a significant increase in the provision of literacy, numeracy and English language learning. The budget for 2002–06 was £3.7 billion, with an increase of spending from £167m in 2001–02 to £995m in 2006–07 (NAO 2008:8). This has enabled the government to fund a major expansion of provision, including enhanced funding to pay for

more hours per learner, extra support and smaller classes. The National Audit Office 2008 report states that not only was in 2004 the public service agreement target achieved of 750,000 learners improving their skills, the Department for Innovation Universities and Skills (DIUS)[2], (formerly the DfES) which is responsible for the strategy, was on course to meet its second target of an additional 1.5m by 2010. So we can say that many of the targets set in 2001 have been realised. And yet while on the surface the picture looks positive, there have been concerns over the implementation of the strategy and the effectiveness of the Skills for Life programme in areas such as the measurement of achievement and the quality of teaching and learning.

The Skills for Life infrastructure

The Department for Education and Skills (DfES) made clear at the outset that the aims of the Skills for Life strategy needed a radical new approach. This was because the basic skills and ESOL sectors had been under-funded for a long time and 'much provision has been ineffective at improving literacy and numeracy skills delivery, with few learners achieving a satisfactory rate of progress' (Skills for Life Strategy 2001:8). This called for the development of 'infrastructure': tools and procedures conceived and commissioned by the DfES. And yet, many of the strategic decisions and approaches that were adopted in the early stages were derived from existing practice. We shall review some of these aspects and their impact in the following sections.

Setting national standards and curricula

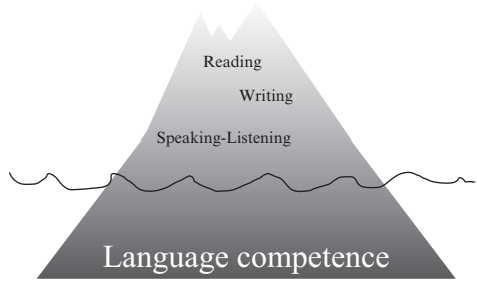
The DfES took the decision in 2000 that literacy and numeracy would have their own standards; but not ESOL. Instead the field should adopt the adult literacy standards for native English speakers.[3] This policy decision was a radical and unique step since, to my knowledge, no other country in the world assesses the skills of first and second language speakers through a single set of standards. Yet despite the fact that the approach favoured by the DfES was so different from practice elsewhere, underpinning evidence of its merits was not supplied nor has its impact been evaluated.

Looking at the use of the national literacy standards from the point of the second language speaker, it is analogous to the position of a British citizen who moves to the Middle East or Italy and who wants to learn Arabic or Italian. This person is asked to work towards and be assessed against standards that apply to native speakers of these languages. This situation is problematic because native speakers can be expected to have fluent language skills, even if they have weaknesses in the way they use them. By contrast, the learning load for second language speakers is quite different. These learners need to develop an understanding of aspects that the first language speaker is

already able to use: the structure and grammar of the language, vocabulary, pronunciation, stress and intonation and the appropriate use of English in the right context. In addition, a priority for the learners is to develop the ability to understand spoken English, by which we mean the ability to identify words in a stream of sound, a process called lexical segmentation. This skill is crucial for the development of listening and other language skills but simply absent from the national literacy standards and the ESOL core curriculum.

The iceberg picture below exemplifies the differences between first and second language learners. People who speak English as their first language need to improve their ability to handle the skills of reading, writing and speaking and listening. By contrast, the priority for second language speakers is to build up their understanding of how English works and to develop their competence in understanding and producing the language.

Figure 2 The nature of English language development



Reading skills

The text below exemplifies how first and second language speakers might tackle the skill of reading. It originates from a bank of national literacy test material which all learners, first and second language speakers, are expected to take to achieve an outcome under Skills for Life. Candidates are expected to read the text and answer five questions. The design of the test is the responsibility of the Qualifications and Curriculum Authority (QCA), the regulatory body for publicly funded qualifications in England.

> Fire officers and police are investigating an explosion that reduced a restaurant and several shops to rubble. One unidentified man was taken to Jubilee Hospital in Park Street after the blast, which involved gas or flammable materials and which is being treated as suspicious.
>
> QCA: *Adult Literacy Test* 2004

While first language speakers may not be fluent readers, if they can decode words (i.e. identify individual letters and assemble these into words), they are very likely to know their meaning, apart from perhaps the word *flammable*. By contrast, a second language speaker may be able to decode a word but not know its meaning. This means that, having made the effort to read, they are no further in understanding the text. This is very likely to happen with the text above, as Leech, Rayson and Wilson's (2001) word frequency list shows. This is based on the 100-million word British National Corpus and provides the following information on the vocabulary in the paragraph above: *investigate* (55 occurrences per 1 million words), *explosion* (22), *reduce* (178), *rubble* (fewer than 10), *identify* (133), *blast* (10), *flammable* (fewer than 10) and *suspicious* (14)/*suspicion* (23). The lower the frequency of the words, the less likely learners are, statistically speaking, to have encountered them. This is particularly relevant for second language speakers who are much more likely not to have encountered words before. In addition, it appears that there is a limit to the ability of second language speakers to deduce the meaning of new words in context. Research by Laufer (1992) and Nation (2001) suggests that most learners find it difficult to infer the meaning of new words, if they know fewer than 95% of the words of a text. Khalifa and Weir (2009:81) suggest an even higher percentage of 97% to enable ease of reading, especially for higher levels of language skills.

Speaking and listening skills

A second decision taken by the DfES was to treat speaking and listening skills as one skill called 'communication', in line with practice applied to the national curriculum for primary and secondary education and for key skills. This means that teachers and learners work on three components: communication, reading and writing. It is not clear why this approach was felt to be appropriate nor does it appear to be based on research evidence. Many young and adult first language speakers who perform below the expected norm in spoken communication show differences in their ability to speak and understand, with their spoken language skills and command of register typically below their ability to understand spoken language. However, the decision to treat the skills of listening and speaking as one entity is particularly problematic for second language learners because there is often a major difference in their level of competence in speaking and listening. It is important to capture the ability of the learners to handle these skills, not only during final assessment but also on entry and while the learners are on the course, as a basis for planning for learning. There is plenty of anecdotal evidence of negative washback of the combined treatment of 'communication' on assessment and teaching. This area would benefit from research.

Benchmarking achievement

Table 2 gives an overview of the DfES' calibration of levels of performance to other frameworks. The first makes a direct link between the expected level of skills of adult learners and those of primary and secondary pupils. This is because the DfES made an explicit link between the strategy for adults to that of young learners: 'Our strategy will build on our literacy and numeracy success in schools' (DfES 2001:8). However, the alignment of the achievement of young learners to that of adult second language learners is problematic, not least since even by the age of five, children can be expected to have learned much, if not all, of the structure of the English language; whereas, as we have seen, the same statement cannot be made of adult second language speakers.

The department subsequently matched the national literacy standards to the *Common European Framework of Reference*, which was produced by the Council of Europe (2001). However, there is uncertainty over the accuracy of their alignment. For example, the performance of language learners indicates that the level 1 of the national literacy standards is well below that of B2. However, they are described as being at the same level in the DfES document *Pathways to Proficiency* (2003a).

Table 2 DfES calibration of performance levels to UK and European frameworks

National adult literacy standards	Age equivalent	Council of Europe levels
	GCSE A–C Achievement at age16	C2
Level 2		CI
Level 1	GCSE D–F Achievement at age 11	B2
Entry 3	Achievement at age 9	B1
Entry 2	Achievement at age 7	A2
Entry 1	Achievement at age 5	A1

Skills for Life: The National Strategy for Improving Adult Literacy and Numeracy 2001:46
DfES: Pathways to Proficiency, London: DfES Publications 2003.

The national standards and their curricula

The national literacy standards and the ESOL and literacy curricula are driven by competence-based performance criteria. Achievement is defined in terms of communication requirements and the settings in which the

communication takes place. It is the function, type of activity and situation which drive the evidence requirements rather than the language used and its appropriateness for the context. This is problematic because the match between function and activity on the one hand and language level on the other is not always as straightforward as it might appear. Many functions can be handled at different language levels, for example one might give personal information using very simple or highly complex language. The variation in output makes it hard to assess when the candidate might have achieved the right standard of language use or indeed what the expected standard should be. To demonstrate this conundrum Table 3 shows three descriptors for Speaking and Listening at Entry 1 and 3; Reading at Entry 3; and their rubrics as they appear in the literacy and ESOL curricula.

Table 3 Comparison of Skill descriptors across ESOL and Literacy curricula

Literacy Standards: Entry 1 *Speak & Listen to Communicate* Descriptor: **Speak clearly to be heard and understood in simple exchanges**	
ESOL curriculum	**Literacy curriculum**
Use stress and intonation to make speech comprehensible to a sympathetic native speaker Examples of language use: Station, computer Can I smoke here? Can I see the manager? (with rising intonation)	Explain a straightforward purpose clearly and appropriately in the context of work, leisure or study, e.g. to the teacher, explaining general aims for a job at the end of the course

National Literacy Standards: Entry 3 *Speak & Listen to Communicate* Descriptor: **Speak clearly to be heard and understood using appropriate clarity, speed and phrasing**	
ESOL curriculum	**Literacy curriculum**
Use stress, intonation and pronunciation to be understood and making meaning clear Examples of language use: He's as tall as his father now The leg's much worse than before	Make a contribution to an informal meeting which is clear, audible and appropriately paced, e.g. giving a vote of thanks to fire officers for visiting neighbourhood fundraising event

National Literacy Standards: Entry 3 *Reading with text focus* Descriptor: **Obtain specific information through detailed reading**	
ESOL curriculum	**Literacy curriculum**
Read an explanation of how something works in detail in order to operate it from a letter or card	Locate and read specific information, e.g. listings in a local newspaper Check details of the date and time of an appointment

These descriptors are problematic on several fronts. It is hard to see how it is possible to assess learner achievement against a statement such as 'Speak clearly to be heard and understood in simple exchanges'. It is equally difficult to differentiate between the very similar descriptors for communication at Entry 1 and Entry 3, even though they are at substantially different levels of performance. A third concern is that, in the examples above, the level of skills expected for the literacy and language curricula is by no means the same. This is not just a localised problem: the ESOL curriculum appears to be set at a much lower level than the literacy curriculum for native English speakers.

In summary, second language speakers experience a learning trajectory which differs in many respects from that of first language speakers, both in the processing of language, cognitive development and the skills to be learned. The national standards and the ESOL curriculum lack a linguistic framework and appropriate instruments to capture the language development of second language speakers against defined milestones. The opaqueness of the standards makes variability of assessment more likely across individual teachers, providers and awarding bodies. Indeed, a comparison of the assessment tasks set by the various awarding bodies indicates that there is substantial variation between them. It is worrying that the national literacy standards and the ESOL curriculum fail to reflect the learners' evolving language skills accurately and that they do not provide sufficiently meaningful benchmarks for teachers. Since, as the Skills for Life strategy (2001) had already acknowledged, much of the provision was ineffective, well-structured standards for ESOL would have made for an excellent start to raise the quality of delivery.

The strategy as educational intervention

Taking the wider context, Tim Oates (2007:144) expresses concern over 'the rapidity of change and turnover in major innovation' in education and training in England. While Oates' examples are drawn from other areas, his concerns apply equally to the introduction of the Skills for Life strategy: the introduction of major innovations without a pedagogic rationale, a lack of consideration of washback on classroom practice, and a lack of safeguards for learners caught up in these innovations. In the case of the Skills for Life strategy, it is not just that new systems were put in place rapidly and without piloting or meaningful consultation, they have been in existence without evaluation. The design of functional skills, the planned successor standards to the national literacy standards, is in danger of repeating the same cycle: evidenced-based assessment practice and methodology from the field of English and foreign language teaching have been ignored; and inaccuracies found in the national literacy and key skills standards are being transferred into the new standards. Responses to consultations and attempts to inform the development of the functional skills standards have been put aside and

there appears to be currently no process whereby draft standards and qualifications can be scrutinised by the practitioner and research communities.

The measurement of government targets

We now turn to the achievement of the Public Service Agreement targets for qualifications, DIUS' main measure of success for the Skills for Life strategy. As we saw earlier, these targets have been met but there is considerable concern over how they have been met as well as the impact that the target setting has had on the provision of literacy, numeracy and language. Most disturbingly perhaps, the National Audit Office has declared 'the data systems underlying this public service agreement target to be not fit for purpose' (NAO 2008: 44). While the National Audit Office reports that action has been taken on many aspects, several issues remain. In the first place, the National Audit Office reports record that in 2001 38% of the basic skills qualifications were achieved by the Skills for Life target group, i.e. people of 19 and older. This percentage had increased to 52% by 2006/07 (NAO 2008:18). The other 48% of the qualifications were gained by students who did not participate in Skills for Life provision: 16–18 year-olds who studied maths and English for GCSE and key skills qualifications. It is of concern that, even in the most recent academic year for which figures have been published, almost half of the qualifications were achieved outside the Skills for Life context. This fact alone makes it hard to defend that the data on achievements count as evidence of the impact of the Skills for Life strategy.

In addition, there are anomalies within the Skills for Life sector itself. There is nothing to stop providers putting learners through the test without a meaningful programme of learning and claiming qualifications if the learner passes. Indeed, there is anecdotal evidence that this happens but its impact is unknown. Secondly, the NAO report reports a sharp drop in the number of passes for the national literacy test achieved for ESOL (2008:18). This may be explained by the fact that many second language speakers who take the national literacy test are registered as literacy rather than language learners. A second contributory cause to this fluctuation in numbers may be the considerable variation in the difficulty of the national test items. Even if the government is content with the data collection to justify Skills for Life expenditure, the manner in which the data are collected provides insufficient information by level of achievement and by the three target groups of literacy, numeracy and language. Nor does the data collection as it stands allow for the evaluation of the educational value of Skills for Life programmes, the allocation of resources and the quality of delivery.

Then there is the national literacy test itself. Despite its title, it only tests a narrow range of reading skills, topics, text length and text types, nor is there any assessment of writing. Since writing is the skill that is most difficult for

most learners – first and second language speakers alike – the candidate's success in passing the national literacy test cannot be said to mean that they have met the standards for both reading and writing.

Conclusion

The Skills for Life strategy has undoubtedly brought benefits for many language learners, not least because they have had access to more language provision than was the case before the strategy was introduced. As to whether it has been *better* provision, evaluations of inspections indicate that ESOL provision has improved over time to a current rating of 'satisfactory' as the most recent OFSTED evaluation indicates (2008:5). However, the report also states: 'the proportion of provision that is good or outstanding has not increased and remains too low' (2008:5). It is my contention that this is at least in part caused by a lack of clarity of what skills and knowledge are to be taught and assessed.

So where next? It seems to me that a review of the Skills for Life strategy is timely. Areas for exploration should be how best to provide the field of ESOL with its own set of standards and a core curriculum which reflect evidence-based research on second language acquisition. The two disciplines of literacy for first language speakers and language learning for second language speakers should be acknowledged as distinct, both when constructing assessment standards for language learning and for teacher training. DIUS should look to harness the data collection on national targets to inform the analysis of the quality and effectiveness of learning programmes. The two major drivers of current educational policy, target setting and funding, need to be balanced by a proper focus on the learner and the effectiveness of the language learning process.

The National Audit Office and the Committee of Public Accounts have made increasingly constructive comment on quantitative measures, especially targets, but government committees have been less successful in the evaluation of qualitative aspects, especially on the quality of teaching and learning, and teacher training.

In the wider arena, politicians and policy makers should find a better balance between the pre-occupations of politics, with its focus on short-term results, and the need to improve the quality and effectiveness of learning over time. With this should come better training for civil servants so that they are in a position to assess the implications and consequences of proposed policy changes. It would also be advisable to keep operational and executive management separate so as to maintain objectivity when reviewing the effectiveness of policy.

The Skills for Life strategy set out to improve the skills of the English-speaking population in England. The numbers of legitimate residents who

access ESOL provision are such that they now form a large part of the constituency of learners. A shift in policy is not only desirable, it is necessary to pay just attention to their needs. It is not just in the interest of other language speakers that they learn English so that they can contribute to society and the economy; it is in the interest of all the citizens of England.

Notes

1. Please note that the United Kingdom consists of four countries: England, Scotland, Wales and Northern Ireland, each with the power to design its own educational policy. Wales and Northern Ireland have largely adopted the Skills for Life approach while Scotland has its own strategy and provision for literacy and language learning. For further detail, see Schellekens (2007).
2. At the time of writing the name of the government department responsible for the Skills for Life strategy is the Department for Innovation Universities and Skills (DIUS). Before 2006 it was called the Department for Education and Skills (DfES). In June 2009 the department was renamed: the Department for Business, Innovation and Skills.
3. This approach had previously been applied to the national curriculum for primary and secondary education and to the key skills standards which are aimed at 16–19 year-olds.

Bibliography

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

DfES (2001) *National Strategy for Improving Adult Literacy and Numeracy Skills,* London: DfES.

DfES (2003a) *Pathways to Proficiency,* London: DfES Publications.

DfES (2003b) *The Skills for Life Survey: A National Needs and Impact Survey of Literacy, Numeracy and ICT Skills,* London DfES (RR490), www.dfes.gov.uk/research

DIUS/DfES (2001) *National Standards and curricula,* London: DIUS/DfES, www.dfes.gov.uk/readwriteplus/curriculum_documents

Ethnic Minority Employment Taskforce (2006/8) *Ethnic Minorities in the Labour Market*, www.emetaskforce.gov.uk

Home Office (2008) *Accession Monitoring Report,* (May 2004 to March 2008), www.ukba.homeoffice.gov.uk/sitecontent/documents/aboutus/reports/accession_monitoring_report/report15/may04mar08.pdf?view=Binary

Horak, T (2008) *Skills for Life Exams,* Joint ESOL/TEA Newsletter Canterbury: IATEFL.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and practice in assessing second language reading*, Cambridge: UCLES/Cambridge University Press.

Laufer, B (1992) How much lexis is necessary for reading comprehension? in Arnaud, P and Béjoint, H (Eds) *Vocabulary and Applied Linguistics*, Macmillan, 126–32.

Leech, G, Rayson, P and Wilson, A (2001) *Word Frequencies in Written and Spoken English,* Harlow: Pearson.
Moser, C (1999) *A Fresh Start: Improving literacy and numeracy,* London: Department for Education and Employment.
Nation, I S P (2001) *Learning Vocabulary in Another Language*, Cambridge: Cambridge University Press.
National Audit Office (2008) *Skills for Life: Progress in Improving Adult Literacy and Numeracy*, London: The Stationery Office.
Oates T (2007) Protecting the Innocent – the Need for Ethical Frameworks within Mass Educational Innovation, in Saunders, L (Ed.) *Educational Research and Policy Making,* London: Routledge Taylor and Francis, 144–174.
Office for National Statistics for access to census data on country of birth by ward, borough and nation for England and Wales, www.ons.gov.uk/census/ and www.nomisweb.co.uk
OFSTED (2008) *ESOL in the Post-compulsory Learning and Skills Sector: An Evaluation,* London: OFSTED.
QCA and DfES (2003) *Pathways to Proficiency*, www.dfes.gov.uk/readwriteplus/bank/PathwaysToProficiency.pdf
Schellekens, P (2001) *English as a Barrier to Employment, Education & Training,* DfES.
Schellekens, P (2007) *Oxford ESOL Handbook,* Oxford: Oxford University Press.
Storkey M, Maguire J, and Lewis R (1997) *Cosmopolitan London,* London: London Research Centre.

7 The requirements of the UK test for citizenship and settlement: critical issues and possible solutions

Szilvia Papp
University of Cambridge ESOL Examinations

Abstract

Nation states have different objectives for introducing tests for citizenship and permanent residency, a major aim being gate-keeping. They also have different views on what the content and language requirements of tests for citizenship and settlement should be. The current UK test for citizenship and settlement requires a minimum proficiency level at Entry 3 within the National Qualifications Framework (NQF, Qualifications and Curriculum Authority, QCA, 2004) or B1 in the Common European Framework of Reference for Languages (CEFR, Council of Europe, CoE, 2001).

This paper examines the Life in the UK test and the publicly available materials provided for study towards it within a framework developed by Kunnan (2008) and a guide for policy-makers developed by members of the Language Assessment for Migration and Integration (LAMI) subgroup within the Association of Language Testers in Europe (ALTE) in collaboration with the Council of Europe (CoE). It investigates the Life in the UK test for qualities of fairness and validity. In particular, the question is raised to what extent and how the language used in the test materials reflects the targeted level of proficiency and language use domain, i.e. functional competence required for the successful demonstration of citizenship and settlement in the UK.

Introduction

This paper examines the Life in the UK test and the publicly available materials provided for study towards it within a framework of test fairness developed by Kunnan (2008). Throughout the evaluation, reference is made to a guide developed by members of the Language Assessment for Migration and Integration (LAMI) subgroup within the Association of Language Testers in Europe (ALTE) in collaboration with the Council of Europe (CoE) (ALTE

LAMI/CoE 2008). The Life in the UK test is evaluated by each criterion demonstrating qualities of fairness and validity and critical issues are discussed and solutions provided as suggested by the ALTE LAMI/CoE (2008) guide.

In particular, the question is raised to what extent and how the language used in the test materials reflects the targeted level of proficiency and language use domain (functional competence required for the successful demonstration of citizenship and settlement). Through corpus analytical methods, it is shown that the test in its current format does not assess the functional competence required for the successful demonstration of citizenship and settlement, i.e. effective participation in everyday social life, employment, and study. The level of English assumed in the test materials is higher than the stated target and the content does not cover the relevant aspects of citizenship and residency.

It is argued that for the test to be a valid and fair test of language proficiency for the purposes of citizenship and settlement, first the construct to be tested would need to be identified through a definition of 'language use for citizenship and settlement'. Then a list of functional descriptions similar to the 'Can Do' statements within the CEFR would need to be developed to inform language test construction. As an alternative form of assessment for migration purposes, the European Language Portfolio (ELP) is proposed, especially those examples of ELPs which were developed with migrant populations in mind, such as the Milestone portfolio developed for teenage and adult immigrants learning the language of the host community in Ireland.

Recent history of language assessment for migration

This section offers a brief overview of the recent history of language assessment for migration purposes in the UK. The precursor of the recent language testing for citizenship and settlement in the UK is the publication of the education White Paper entitled *Excellence in Schools* in 1997, when Citizenship became a separate subject rather than a cross-curricular theme in schools in England. Following this, the Department for Education and Skills (DfES) introduced Citizenship education into secondary schools in 2002. In the same year, the Nationality, Immigration and Asylum Act foreshadowed the latest round of activities in the third millennium around migration (see e.g. HO 2007a, b, c). As Saville (2008) points out, these renewed activities are strongly reminiscent of concerns and solutions of the past 100 years.

The Life in the UK test

Since July 2004, applicants for UK citizenship have had to demonstrate knowledge of the English language. In June 2005, a test with a language

element, called the Life in the UK test was developed by UfI/Learndirect Ltd for the Home Office (HO) and piloted at eight test centres. All applicants for citizenship or British nationality have had to pass this test since 1 November 2005. In addition, since 2 April 2007, applicants for settlement have had to take the test to be granted indefinite leave to remain. There is no sample test available[1], but the HO offers guidelines on their website about the content and format of the test in an online tutorial. The materials provided by the HO towards preparation for the test include a Handbook entitled *Life in the UK: A journey to citizenship*. The original Handbook (HO 2004) was published in December 2004 for teachers and mentors of immigrants. A revised and simplified version of the Handbook (HO 2007) was published specifically for candidates in March 2007. The targeted proficiency level of the test is ESOL Entry Level 3 or above in NQF terms, or B1 or above in CEFR terms. For those candidates not at this level yet, learning materials for ESOL with citizenship courses were developed by the National Institute for Adult and Continuing Education (NIACE) and LLU+ (formerly the London Language and Literacy Unit) and piloted with 18 ESOL providers between September 2004 and March 2005, prior to the first launch of the test (cf. Sunderland and Taylor 2008).

Current exemptions from the UK test for citizenship and settlement include criteria such as age (those under 18 or over 65 years old are exempt), disability (individuals are exempt who are prevented from learning English, either because they are suffering from long term illness, or because they have a permanent disability which severely restricts mobility and ability to attend language classes, or are suffering from a mental impairment rendering them unable to learn another language), and special needs. Even if special arrangements need to be made to learn English or take the test, candidates still have to meet the requirements. Test accommodations, such as extended time, and arrangements for candidates with limited mobility or visual impairment are in place.

By making language proficiency part of the requirements in order to gain citizenship and settlement rights and carry out responsibilities as citizens or permanent residents in the country, there is an underlying assumption that proficiency in the language of communication in the wider society is a democratic responsibility and a democratic right. This clearly has implications not only for the candidates, but also for wider society, such as policy makers, government, employers, colleges, and schools. Also, it has implications for the development and administration of tests conceived and designed for these purposes, that is, for professional testers and stakeholders. Tests developed and accepted for the purposes of migration, citizenship and settlement, which are very high stakes for all concerned, need to be defensible. These tests need to be evaluated, and a strong, theoretically based, empirically supported and logically developed argument needs to be presented in their defence.

The specific testing context and purpose always determines the appropriacy

of the framework in which a test can be evaluated. Since one of the concerns around tests for migration purposes is test validity and hence quality, the chosen framework could be Weir's (2005) socio-cognitive framework for test validation which focuses on the social context of the testing situation and cognitive aspects of the candidates and test demands (for a critique of the Life in the UK test within the socio-cognitive framework, see Papp and Wright 2006). Alternatively, if the primary concern is fairness, Kunnan's (2008) framework of test fairness could be chosen which focuses on qualities of test fairness and ethics.

In this paper I address the question whether there is a defensible argument for the current Life in the UK test, especially but not exclusively, its language requirement, in terms of Kunnan's test fairness qualities. Members of the Language Assessment for Migration and Integration (LAMI) subgroup within the Association of Language Testers in Europe (ALTE) have been working closely with the Council of Europe (CoE) and relevant UK government departments, such as the Home Office Border and Immigration Agency (HO BIA) and the Advisory Board on Naturalisation and Integration (ABNI) since 2002, the year of the introduction of the policies on assessment for migration purposes (cf. Saville and van Avermaet 2008). The ALTE LAMI group has produced an outline for policy makers to enable them to better judge the merits of language tests for social cohesion, citizenship and settlement, a guide henceforward referred to as ALTE LAMI/CoE (2008). The issues that arise in the evaluation of the Life in the UK test will be discussed and informed by considerations set out in this guide.

Evaluating the Life in the UK test

Kunnan (2008) defines an ethical and fair test as displaying a) comparable or equitable treatment in the testing process, b) comparability or equality in outcomes of learning and opportunity to learn, c) absence of bias in test content, language and response patterns, and d) comparability in selection (and prediction). These are requirements that the Life in the UK test should aim to fulfil, since 'test fairness is relevant to all types of language test and for all target candidates, but is especially important in the case of tests of language for migration, residency and citizenship, due to the serious implications for the test taker in terms of civil and human rights' (ALTE LAMI/CoE 2008).

For the Life in the UK test, a defensible argument is needed, backed by evidence collected from Kunnan's (2008) **five test fairness qualities**:

1. Validity
2. Absence of bias
3. Access

4. Administration
5. Social consequences

According to the ALTE LAMI/CoE (2008) guide, the following aspects need to be considered in order to evaluate any test developed and used for migration purposes:

1. Test purpose and real-world demands on test takers
2. Linguistic demands
3. Appropriate level of difficulty linked to the CEFR
4. Test specifications (covering test format, task and item types, assessment criteria, item writing, pretesting, test administration, special requirements, marking and grading)
5. Monitoring (examiners, candidate responses and demographic information in order to be able to equate test versions and identify possible bias and to ensure test functionality and quality).

All of these aspects relate to and exemplify the five fairness criteria set out by Kunnan (2008) above. For the criterion of **validity**, the following questions need to be answered in the case of the Life in the UK test:

- For what purpose(s) was the test developed?
- What does the current test (both in its first and second forms) measure (i.e. the construct behind the test)?
- Who has been involved in developing, evaluating and revising the test?
- How best can language skills needed for integration and social cohesion be tested?

According to the ALTE LAMI guide, it is imperative to first decide on and explicitly state the purpose of the test.[2] The authors suggest that policy makers need to reflect on whether the primary objective of a test for migration is:

> 1. motivating learners (to help them use and improve their current competence in the target language),
> 2. ascertaining whether their competence is sufficient for participation in well-defined social situations (e.g. study or work and also other social situations more connected with the exercise of citizenship),
> 3. making decisions which affect their legal [as well as human and civil] rights, such as their right to remain in a country or acquire citizenship of it (ALTE LAMI/CoE 2008).

Language tests developed for migration purposes should primarily fulfil the second purpose. The other two are applied objectives for test use and interpretation of test results. All of these purposes have been articulated and debated in the public and professional discourse in the UK and beyond. Language testing professionals can and should advise on the second purpose, since they are well placed to do so with authority and competence, based on

their expertise and professional codes. For the other two applied objectives impact and washback studies are needed in order to ascertain the beneficial and potentially harmful effects of a test on society as a whole and its institutions, and teaching and learning within the classroom.

The ALTE LAMI guide points out that 'only when the purpose has been clearly defined is it possible to identify the real-world demands that test-takers will face (e.g. the need to take part in societal processes and exercise the rights and responsibilities of citizenship) and which should be reflected in the test' (ALTE LAMI/CoE 2008). As well as informing test development, a clear and explicit purpose helps to clarify test takers' expectations towards test type, test content and marking criteria. All of these aspects need to be specified and made public in the test specifications. This will contribute to test fairness and will also allow other members of society to interpret and use the test results appropriately.

Unfortunately, there has been a considerable amount of criticism levelled against the current Life in the UK test since its conception and introduction. A major indictment can be found in Lord Goldsmith's recent Review of Citizenship report (March 2008). As far as the purpose of the Life in the UK test is concerned, Lord Goldsmith notes that 'the present test is not seen typically as a stimulus for learning, though that was one of its stated aims' (Lord Goldsmith 2008:118). It is clear that policy makers need to reflect on the primary purpose of the Life in the UK test, and clearly articulate it to the stakeholders. This would directly impact on the test specifications, inform test revision, and help in the interpretation of test results.

Once the primary purpose is established, the test specifications need to be published. Test specifications include an explanation of the process of establishing the needs of target candidates, usually termed *needs analysis*. In the needs analysis exercise, it is very important to heed the advice of the authors of the ALTE LAMI/CoE guide:

> When conducting this needs analysis, it is also necessary to take into account the fact that there are various subgroups of migrants with their own specific needs. Those, for instance, who want to join the job market as soon as possible often have different needs from those who are planning to raise young children at home. In a needs analysis, it is good practice for language test developers to define the relevant contexts and situations and other characteristics of the target group. In planning such needs analyses, policy makers should be sure to set aside sufficient resources and delegates from different parts of society should be involved in the definition of the needs. Language tests for study and work are, in most cases, taken by groups of candidates which are homogeneous with respect to educational background and cognitive skills, whereas tests for integration and citizenship (tests to acquire civil rights) must cater for a full range of possible candidates, and must therefore be accessible to

> people with little or no literacy skills, as well as those with a high level of academic education (ALTE LAMI/CoE 2008).

Therefore, it is advisable to carry out a thorough analysis of test taker characteristics and their specific needs in order for the Life in the UK test to be better targeted and be fit for its stated purpose. This is crucial in order to be better able to define the construct (knowledge, skills and abilities) that lies behind tests developed for these purposes: language knowledge, skills and abilities needed for citizenship and residency in the UK.

This step is clearly important, even though it can be argued that any needs analysis tends to be a taxonomic exercise which typically restricts the construct underlying a test. In defining the construct, some questions policy makers and appointed test developers need to address include the following:

- What level of social participation are migrants (would-be citizens and settled residents) expected to aim for?
- What level of 'exercising their civil rights and responsibilities' are they expected to demonstrate – up to representing other people in democratic processes?

While considering the above questions, the specifications set out for language tests for migration, settlement and citizenship should identify minimum requirements set for candidates. Hence, the questions to be answered are:

- What minimum language competence do individuals need to be able to exercise their rights and responsibilities and participate in social situations?
- In what ways do we wish candidates to exhibit the required level of functional competence for successful demonstration of knowledge and skills for citizenship and settlement?

As the quote from the ALTE LAMI/CoE guide above shows, it is also important to consider the team involved in developing a test for migration purposes. Members of the advisory board ABNI who advised on the development of the test included 'experts in the fields of ESOL, citizenship training, employment of migrants, and community development and integration' (ABNI 2006, Section 2). The Citizenship Test subgroup included 'members [who] have expertise in citizenship and ESOL as well as regional expertise'. The subgroup had the following terms of reference:

> **1** To advise on the evaluation of
>
> (i) the test items
> (ii) the pilot tests (as whole tests), in accordance with basic test assessment principles to ensure that the tests are both valid and reliable and at an appropriate level of 'difficulty'.

> **2** To advise on the piloting of the test on an appropriate sample population.
> **3** To advise on the implementation and administration of the test.
> **4** To advise on the evaluation of the test once in place.
> (ABNI 2006, Section 6)

The item bank for the Life in the UK test was developed between February and May 2005. According to the first annual report of ABNI (ABNI 2006), the advisory group devised, scrutinised and approved items 'based on guidelines regarding question type, language level, and weighting for the text in chapters 2, 3 and 4 of the Life in the UK Handbook' (ABNI 2006, Section 6). The test was piloted on 'candidates from both the target group as well as native English speakers', and 'Ufi was responsible for the trialling and piloting of these test items, and the production of an evaluation report in August 2005. 245 candidates from both the target group and from groups of native English speakers born in the UK and elsewhere completed pilot tests at 8 test centres after reading the designated chapters 2, 3, and 4 of the Life in the UK Handbook, with a mean score of 16.4 / 24, with the pass mark calibrated at 16, resulting in 73% of candidates achieving a pass' (ABNI 2006, Section 6). At the launch of the citizenship test in November 2005, '250 colleagues involved in ESOL delivery, social integration and community cohesion, refugee and immigrant integration, citizenship and the home affairs media' were present to take up 'the opportunity for dialogue' (ABNI 2006, Section 6). Apart from ESOL expertise, it seems that no assessment expertise in general or language testing expertise in particular was involved in the development and evaluation of the Life in the UK test to advise on language assessment-related matters.

As far as the second criterion of **absence of bias** is concerned, at least three checks need to be made about the following aspects of the Life in the UK test:

> a) content or language (Is the content and level appropriate and relevant for candidates?)
> b) the standard (What is the criterion measure and the resulting selection decisions?)
> c) disparate impact on different groups.
> (Kunnan 2008)

According to ABNI, the Life in the UK test is 'a test of knowledge only, and does not require interpretation, analysis or evaluation' (ABNI 2006, Section 4). The **content** of a test for migration purposes should cover knowledge relevant for active participation in societal processes and situations and it should enable users of that knowledge to exercise their civil rights and responsibilities. As the ALTE LAMI guide states, 'once these real-world demands have

been identified, they must be translated into linguistic requirements specifying not only the knowledge and skills, but also the ability level for each that the test-taker is likely to need' (ALTE LAMI/CoE 2008). There is a difficulty in this task, which is fully acknowledged by the LAMI group:

> . . . deriving linguistic requirements from relevant real-world tasks is far less straightforward in the case of migrants and candidates for citizenship. The relation between language proficiency in the official language(s) and the ability to integrate into society and/or exercise the rights and responsibilities of citizenship is looser and far more difficult to pin down. After all, if language proficiency were the only factor in play, all native inhabitants of a country would be fully integrated citizens (ALTE LAMI/CoE 2008).

What this means is that native speakers have the *potential* to be fully integrated citizens. This is because native speakers, even though having a very similar underlying knowledge of the linguistic system of their native language, do not have identical functional performance abilities. What is important in this point is that language is only one of many factors in successful social integration and participation, most of which are still poorly understood. As the ALTE LAMI guide states: 'As this is not the case [i.e. that all native inhabitants of a country are fully integrated citizens], it can be deduced that other factors are also important. The task for the language test developer is, nevertheless, to identify the relevant linguistic demands that apply' (ALTE LAMI/CoE 2008), but also to be aware what other factors impact on these linguistic demands, one of the major ones being literacy.

As far as test **content** is concerned, Lord Goldsmith's (2008) Citizenship Review suggested that the 'ESOL curriculum must accompany a progression route [. . .] from being able to function on a day-to-day basis, through accessing employment, handling bureaucracies, good parenting and educational and career progression' (Lord Goldsmith 2008:113). The curriculum should ideally be reflected in the Life in the UK test, which in its current format entirely lacks these aspects.

In the Life in the UK test, the **language** criterion is set at ESOL Entry Level 3 or B1 of the CEFR. 'The current level was arrived at after extensive consultations with the educational sector, and with NGOs representing and working with immigrants and refugees' (ABNI 2006, Section 3).

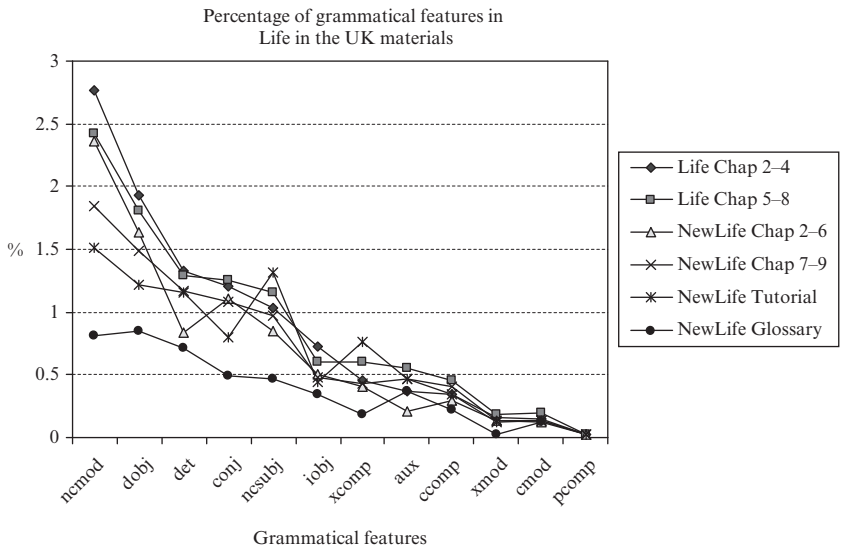
To investigate the level of language competence required in Life in the UK test, corpus methodologies were used to study some linguistic features in the published test materials. Figure 1 shows data separated into materials from the old Life in the UK Handbook published in December 2004 for teachers and mentors of immigrants (Chapters 2–4 which were covered by the test until 1 April 2007 and Chapters 5–8 which were not covered by the test), and the new Life in the UK Handbook published in March 2007 (Chapters 2–6

which are covered by the test since 2 April 2007 and Chapters 7–9 which are not), as well as the Tutorial provided online and the Glossary at the end of the new Handbook.

The materials were tagged and parsed by software developed at the Research Centre for English and Applied Linguistics at the University of Cambridge[3] (see Briscoe 2006). The tagged and parsed text files were run through WordSmith Tools Version 4 (Scott 2004) to find frequencies and percentages of occurrence of each part of speech and grammatical tag for evidence of language level.

Figure 1 Analysis of grammatical features in the Life in the UK materials

Legend:
ncmod = non-clausal modifiers and their heads (the old man in the barn)
dobj = relation btw verbal or prepositional head and the head of the NP to its immediate right (she gave it to Kim)
det = for determiners, including articles, quantifiers, partitives (some men)
conj = for coordinating conjunctions (oranges and clementines or satsumas)
ncsubj = relations btw non-clausal subjects (NPs, PPs) and their verbal heads (the upset man)
iobj = relation btw head and preposition of a PP argument when the PP complement is a NP (flew to Paris from Geneva)
xcomp = relation btw a head and a VP complement (thought of leaving)
aux = for auxiliaries (has been sleeping)
ccomp = relation btw a head and the head of a clausal complement (asked about him playing rugby)
xmod = predicative relations btw modifiers (VPs, APs) and heads (who to talk to)
cmod = relations btw clausal (S) modifiers and heads (although he came, Kim left)
pcomp = relation btw a head and the preposition of a PP argument when the PP complement is itself a PP (climbed through into the attic)



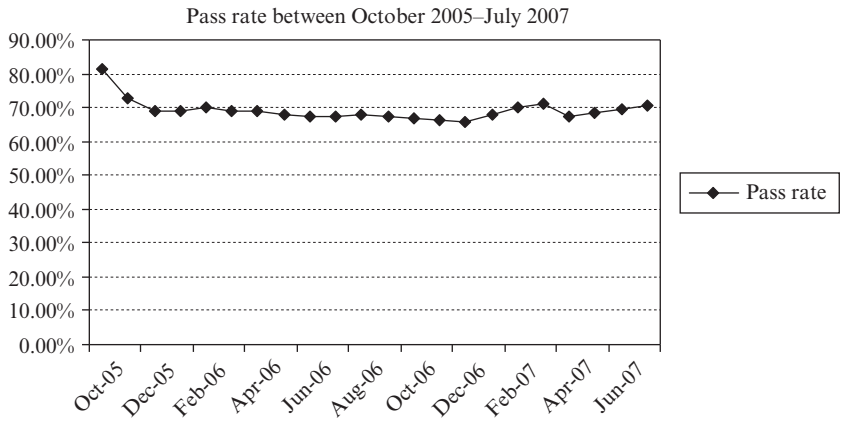
It can be seen in Figure 1 that the effort to simplify the materials provided for learning and teaching towards the test in the new Handbook as well as the online Tutorial and Glossary have had an effect on the percentage of grammatical features in most of the linguistic categories. The question is, whether the simplification has been successful to approximate the targeted proficiency level (B1 in the CEFR).

In relation to this, the ALTE LAMI document gives advice on using the CEFR as a framework of reference for language tests for migration, particularly to enable determining the target level of the tests. Some have argued that the CEFR may not be appropriate, suitable, achievable, or justified for migration purposes (e.g. Alderson 2007, Hulstijn 2007, Krumm 2007). However, in line with the ethos of the CEFR (see e.g. North 2007), the task for policy makers and their appointed test developers is to adapt the framework for the relevant groups of learners and the specific purposes of migration, citizenship and settlement. In the case of language assessment for migration purposes, the relevant groups are similar to naturalistic adult and child second language learners studied longitudinally in second language acquisition studies. Some of the findings and implications of the early and highly influential studies carried out in the first half of the 1990s in Germany with Gastarbeiter groups from different language backgrounds are particularly relevant here (Klein and Perdue 1992). Descriptors for these learner groups' language performance would need to be developed and expressed in language-specific, functional, linguistic, and socio-cultural exponents. These descriptors need to be based on second language learnability considerations within the relevant social and educational context, keeping in mind the characteristic features of the target groups as well as the actual language needs and language use of these learners. For an example of successful application of the CEFR to adult refugee immigrants, young migrants and minority groups in schools in Ireland, policy makers could refer to Little (2007). Of special relevance to this context are examples of the European Language Portfolio developed specifically for these target groups, such as the Milestone portfolio developed in Ireland (Council of Europe 2003).

Policy makers should reflect on the language criterion set for the Life in the UK test, especially in light of the **standard** that is currently set in the test. As Figure 2 shows, 30% of the candidates fail the Life in the UK test (ABNI 2006, 2008). For a very high-stakes test, this rate of failure may not be acceptable or defensible. ABNI originally recommended in their first report 'that a reasonable success rate be in the proximity of 75%', and proposed 'a review of the test if it is consistently below that figure' (ABNI 2006, Section 6). Even the 75% success rate leaves one in every four candidates with a failure as a result of taking the test.

Some analysis has been carried out by the HO to investigate the failure rate in the test: 'countries producing a high number of test applicants yet with

Figure 2 Pass rates between October 2005–July 2007 in the Life in the UK test, based on HO/ABNI data published in ABNI (2006, Section 6)



a pass rate below 50% include Afghanistan, Angola, Bangladesh, Sri Lanka and Turkey' (ABNI 2007, 2008). Apart from these country statistics showing a clear **disparate impact on different L1 groups**, no further data on age, gender, educational background, profession, etc. has been published by HO/ABNI. These are variables test developers and psychometricians usually collect to investigate group differences in test performance. At the same time, these variables are possible causes of group differences in participation in societal processes and situations, and in the exercise of civil rights and responsibilities. Therefore they should also be investigated in an evaluation and revision of the Life in the UK test.

Combining considerations about the purpose of the Life in the UK test, its content and underlying construct, as well as its fairness, Lord Goldsmith has advised the following in his report:

> Government should give consideration to revising the test. Most people born in the UK would struggle to pass the current test and this creates a deep impression of unfairness among people who have to take the test. That in turn affects their willingness to treat the test as part of a learning journey – which must be the underlying objective here – and undermines the credibility of the process (Lord Goldsmith 2008:119).

With regard to the criteria of **access** and **administration**, the following considerations need to be made:

- Do candidates have equal opportunities?
- Do they have comparable educational, financial, geographical, personal access to the course and the test itself?

- Do they have equal opportunity to learn, to practise language learned in everyday life, to become familiar with the test, based on time engaged in high quality learning activities?
- Is the course and the test affordable, are the accepted payment methods accessible?
- Are students able to attend classes and the test and not hindered by geographical distance?
- Are there accommodations for special needs?
- Is familiarity with test conditions, procedures and equipment assumed?

In the case of the Life in the UK test, there are concerns about most of the questions related to access and administration: the availability and quality of ESOL opportunities, inadequate provisions for teaching and learning, long waiting lists for ESOL with citizenship classes, increased fees, some candidates unable to attend classes or take the test due to long working hours or caring responsibilities, insufficient considerations of special needs (including factors such as level of prior formal education, learning skills and literacy, mental health issues, physical impairment) and concerns about using a computer as medium of testing.

In accordance with the Special Educational Needs and Disability Act (OPSI 2001) and Disability Discrimination Act (Directgov 2005), disabled candidates need to be given equal opportunities for learning and assessment. However, the HO requirements to take the exam still hold 'if candidates suffer from a mental or physical ailment that could respond to treatment or therapy'. The HO's list of **test accommodations** does not include accommodations for special needs other than those usually catered for in public exams (deafness, blindness, mobility difficulties). According to the ALTE LAMI guide, special needs in the context of language tests for migration purposes 'may include temporary or long-term physical, mental or emotional impairments or disabilities, learning disorders, temporary or long-term illness, illiteracy in the L1 or target language, regulations related to religion, penal confinement or any other circumstances which would make it difficult or impossible for a candidate to take the test in the same way as anyone else' (ALTE LAMI/CoE 2008). This list could arguably be extended, for instance, to lack of testing experience, lack of funds to cover the costs of a language course and the test, lack of opportunity for familiarisation with the test, and so on. It is clear that this list is different from the usual list of special circumstances, such as blindness or deafness, conditions which make it impossible for people to take a language course or a test designed for sighted and hearing people. Accommodations for special needs seem to be the backbone of fair and ethical language testing for migration purposes, and therefore should receive close consideration in the case of the Life in the UK test.

Lastly, but perhaps most importantly, policy makers and test designers

should consider the **social consequences** of introducing or using a language test for migration and social integration purposes, such as the Life in the UK test. The questions to ask include the following:

- Have the original policy objectives been met?
- Have the arrangements for learning and teaching for the test had the desired consequences, i.e. have they contributed to social integration and cohesion, and civic participation?
- Has the test been evaluated for margin of error and possible bias inherent in all tests?

It is imperative that the margin of error and possible bias inherent in all tests is routinely checked and reduced as much as possible in any test, but especially in tests used for the purposes of migration and social integration. Testers are responsible for making sure that the margin of error and bias are within acceptable limits according to internationally accepted standards listed in ALTE LAMI/CoE 2008, as well as Kunnan 2008, McNamara and Roever 2006. The ALTE LAMI authors caution that:

> However, even though the test may perform perfectly well for a large group of candidates, it cannot take into account each candidate's individual personality traits, learning history and personal history. Thus, seen from the perspective of an individual test taker, scores may not always represent their true ability and it is not possible to exclude all element of test bias with total certainty. Therefore, the overall benefit which it is hoped will be attained by administering the test has to be considered in relation to the consequences of failure in the test, since some of the decisions based on test results may be mistaken (ALTE LAMI/CoE 2008).

As an alternative to testing, policy makers may wish to consider using other means of assessing knowledge and skills needed for migration purposes. As mentioned above, a particularly useful and beneficial way of assessing functional competence for integration and social cohesion are variants of the European Language Portfolio (part of the toolkit of the CEFR) developed specifically for relevant migrant groups, that is, would-be citizens and permanent residents.

Conclusion

The Language Assessment for Migration and Integration (LAMI) subgroup within the Association of Language Testers in Europe (ALTE) has been in existence since 2002, the year when the policy of language testing for migration and social integration purposes started to be formulated in the UK (Saville 2006). The group has been working closely with the Council of Europe and

seeking to contribute to decisions made by policy makers in the relevant UK government departments, such as the Home Office Border and Immigration Agency and the Advisory Board on Naturalisation and Integration.

Policy-making around language assessment for migration is an area that has been constantly changing in response to rapid and unexpected social and economic changes. Since all policy-making decisions on assessment for migration purposes have immediate social consequences on individuals, institutions and society as a whole, it is very important that the initial dialogue and collaboration between professional policy makers, language testers and other stakeholders is maintained and enhanced in order to achieve successful social cohesion and integration in the UK.

It is heartening to see that the UK government has recently been considering alternative ways of language assessment for migration purposes. For instance, for the purposes of access/entry to the UK, a new points based system was phased in in 2008 (HO BIA October 2007, HO UK BA June 2008, and other documents on the HO UK BA website). By accepting more types and a wider range of evidence of knowledge of English and by differentiating specific groups of test takers, the points based system will hopefully act as an instrument for recognition and inclusion rather than a systematic means for exclusion. It will tease apart content (such as knowledge of life in the UK which the current Life in the UK test is based on) from language skills, and address the minimum language requirements for different purposes (entry/access, residency/settlement, citizenship) and for different migrant groups. In the future, decision makers might even consider a system that is able to profile partial competencies and recognise multicompetence which makes up migrants' linguistic capital similar to what is recognised in the European Language Portfolio. If such a system is implemented correctly and fairly, it could ultimately act as a strategic, forward-looking, overarching system for social cohesion (protection, integration, inclusion), rather than being part of the current retrospective management of migration (risk management, 'fire fighting', and inevitably, exclusion) in the UK.

Whatever policy making system is chosen, however, it is very important that only tests that meet rigorous standards should be approved for these purposes. It is here that the work of such language testing professionals as Anthony Kunnan and the LAMI subgroup within ALTE has an important role to play.

Notes

1. Since the writing of this paper the HO has published an official Study Guide (TSO 2008) and a Practice Questions and Answers (TSO 2009) booklet.
2. 'The first step in this process is the precise and unambiguous identification of the purpose of the test. After this is done, principled ways to determine the content and difficulty follow. Finally, the test specifications document,

a document essential in later stages of test construction and review, must be developed' (ALTE LAMI/CoE 2008).

3. I wish to acknowledge and thank Dr Paula Buttery from RCEAL Cambridge for her help in tagging and parsing the Life in the UK test materials.

References

ABNI (Nov 2006) Progress towards integration: 2005/2006 annual report. November 2004–April 2006, <http://www.abni.org.uk/publications/index.html> (accessed 4 February 2009).

ABNI (Oct 2007) Second annual report. April 2006–October 2007, <http://www.abni.org.uk/publications/index.html> (accessed 4 February 2009).

ABNI (Nov 2008) Final report. November 2004–November 2008, <http://www.abni.org.uk/publications/index.html> (accessed 4 February 2009).

Alderson, J C (2007) The CEFR and the need for more research, *The Modern Language Journal* 91, 659–663.

ALTE LAMI Authoring Group/CoE (2008) *Language tests for social cohesion and citizenship – an outline for policymakers*, Council of Europe, <http://www.coe.int/t/dg4/linguistic/MigrantsSemin08_ListDocs_EN.asp#TopOfPage> (accessed 2 February 2009).

Briscoe, T (2006) An introduction to tag sequence grammars and the RASP system parser, *Technical Report* 662, University of Cambridge Computer Laboratory, <http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-662.pdf> (accessed 29 August 2008).

CoE (Council of Europe) (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

CoE (Council of Europe) (2003) Milestone European Language Portfolio, <http://www.eu-milestone.de/files/sites/eumilestone/elp_engl._validated_2003.pdf> (accessed 17 June 2009).

CoE (Council of Europe) (nd) *European Language Portfolio*, retrieved from: www.coe.int/portfolio (accessed 17 June 2009).

DfES (Department for Education and Skills) (2003) *Pathways to proficiency: The alignment of language proficiency scales for assessing competence in English language*, <http://www.desf.gov.uk/readwriteplus/LearninginfrastructureAccreditation> (accessed 17 June 2009).

Directgov (2005) The Disability Discrimination Act (DDA). <http://www.opsi.gov.uk/acts/acts2005/ukpga_20050013_en_1> (accessed 4 February 2009).

HO (The Home Office Life in the UK Advisory Board) (2004) *Life in the United Kingdom: A Journey to Citizenship Handbook*, TSO (The Stationery Office).

HO (The Home Office Life in the UK Advisory Board) (2007) *Life in the United Kingdom: A Journey to Citizenship Handbook*, 2nd ed., TSO (The Stationery Office).

HO (March 2007a) Results of the public consultation on proposals for a Migration Advisory Committee, <http://www.ukba.homeoffice.gov.uk/sitecontent/documents/aboutus/consultations/closedconsultations/establisingamac/> (accessed 4 February 2009).

HO (March 2007b) *Fair, effective, transparent and trusted: rebuilding confidence in our immigration system. An independent and transparent assessment of immigration*. Policy statement, < http://www.homeoffice.gov.uk/documents/ind-review-250706/ind-review-eng?view=Binary> (accessed 4 February 2009).

HO (March 2007c) *Securing the UK border: Our vision and strategy for the future*, <http://www.homeoffice.gov.uk/documents/securing-the-border> (accessed 4 February 2009).
HO BIA (Oct 2007) *Points based system: Procedure for inclusion on the list of approved English language tests*. Tier 1, <http://www.ukba.homeoffice.gov.uk/sitecontent/documents/managingourborders/pbsdocs/> (accessed 4 February 2009).
HO UK BA (June 2008) *Points based system: Procedure for inclusion on the list of approved English language tests*. Tier 2, <http://www.ukba.homeoffice.gov.uk/sitecontent/documents/managingourborders/pbsdocs/> (accessed 4 February 2009).
HO UK BA (Aug 2008) *Points Based System – key documents*, http://www.ukba.homeoffice.gov.uk/sitecontent/documents/managingourborders/pbsdocs/> (accessed 4 February 2009).
Hulstijn, J H (2007) The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency, *The Modern Language Journal* 91, 663–667.
Klein, W and Perdue, C (1992) *Utterance Structure (Developing grammars again)*, Amsterdam: John Benjamins.
Krumm, H-J (2007) Profiles instead of levels: The CEFR and its (ab)uses in the context of migration, *The Modern Language Journal* 91, 667–669.
Kunnan, A J (2008) Towards a model of test evaluation: Using test fairness and the test context frameworks, in Taylor, L and Weir, C J (Eds) *Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity*, Studies in Language Testing Vol. 27. Cambridge: UCLES/Cambridge University Press, 229–251.
Little, D (2007) The CEFR for Languages: Perspectives on the making of supranational language education policy, *The Modern Language Journal* 91, 645–655.
Lord Goldsmith QC (March 2008) Citizenship Review. Citizenship: Our Common Bond, <http://www.justice.gov.uk/docs/citizenship-report-full.pdf>.
McNamara, T and Roever, C (2006) *Language Testing: The social dimension*, Oxford: Blackwell.
North, B (2007) The CEFR illustrative descriptor scales, *The Modern Language Journal* 91, 656–659.
OPSI (Office of Public Sector Administration) (2001) Special Educational Needs and Disability Act (SENDA), <http://www.opsi.gov.uk/ACTS/acts2001/ukpga_20010010_en_1> (accessed 29 Aug 2008).
Papp, S and Wright, S (2006) *Language tests and citizenship: A legacy of the past*, Manuscript, University of Portsmouth, School of Languages and Area Studies.
QCA (Qualifications and Curriculum Authority) (2004) *National Qualifications Framework*, <http://www.qca.org.uk/> (acessed 17 June 2009).
Saville, N (2006) Language testing for migration and citizenship, *Research Notes* 25, 2–4, Cambridge: Cambridge ESOL.
Saville, N (2008) *UK Migration – 100 Years on*, Manuscript, Cambridge: Cambridge ESOL.
Saville, N (2009) Language assessment in the management of international migration: A framework for considering the isues, *Language Assessment Quarterly* 6 (1), 17–29.

Saville, N and van Avermaet, P (2008) Language testing for migration and citizenship, in Taylor, L and Weir, C J (Eds) *Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity*, Cambridge: UCLES/Cambridge University Press, 265–275.

Scott, M (2004) *WordSmith Tools Version 4*, Oxford: Oxford University Press.

Sunderland, H and Taylor, C (2008) Citizenship materials for ESOL learners in the UK, in Taylor, L and Weir, C J (Eds) *Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity*, Cambridge: UCLES/Cambridge University Press, 310–320.

TSO (The Stationery Office) (2008) *Life in the United Kingdom: Official Citizenship Test Study Guide*, TSO (The Stationery Office).

TSO (The Stationery Office) (2009) *Passing the Life in the UK Test: Official Practice Questions and Answers*, TSO (The Stationery Office).

Weir, C J (2005) *Language Testing and Validation: An evidence-based approach*, Basingstoke: Palgrave Macmillan.

Section Two
Insights on testing within language teaching and learning

8 Setting language standards for teaching and assessment: a matter of principle, politics, or prejudice?

Lynda Taylor
Consultant, University of Cambridge ESOL Examinations

Abstract

In language education linguistic variation and diversity raise both theoretical and practical issues about *what to teach* – in relation to pedagogy, materials and training, and *what to test* – in terms of the standards, norms, models and judgement criteria to be adopted. Decisions may be influenced by socio-political sensitivities, even prejudices, about whose language should be the focus of attention, as well as by the ever increasing pace of language change in today's fast-moving world which is so strongly shaped by modern information and communications technology. Debate over the relative merits of 'native speaker' and 'non-native speaker' models is symptomatic of the dilemma. Language testers need to be able to account for language variation within the model of linguistic or communicative competence underpinning their tests and to understand how it can affect the validity, reliability, practicality and impact of the tests they offer.

This paper begins by considering the nature of standards and the role that the 'setting of standards' and 'standardisation' play in the business of assessment. It then describes aspects of linguistic variation and considers some implications these have for language teaching, learning and assessment, especially for testing agencies. It touches upon the potential factors that shape policy and practice – including factors such as politics and prejudice. The final part of the paper explores how far two frameworks of reference that have recently been proposed within the language testing and assessment community for test validation and evaluation purposes might help us in constructing a principled and pragmatic approach to setting language standards for teaching and assessment.

Introduction

> *'It is a truism that modern life runs by the clock. Clocks synchronize our communal activities, and that they do so is at once useful and tyrannical.'*
> (Stephens 1989:1)

We are familiar with the idea of setting and accepting regional or international standards in many aspects of society and everyday life. We are used to working with standards relating to weights and measures, temperature, even time itself. Indeed it was a process of 'time standardisation' during the 19th century that enabled delegates from all over the world to reach Cambridge for the ALTE 3rd International Conference in April 2008. The development of national communication networks – especially the 19th century railway and telegraphic systems in Britain and the US – was instrumental in the emergence of a uniform standard of time; this new *standard time* allowed some co-ordination at a supra-local level, where previously *local time* had been perfectly adequate even if it differed somewhat from the local times of other localities near and far. Efforts to synchronise different communities with one another were essential to ensure that travel connections could be made, as well as to prevent accidents between trains using a shared rail network.

As international travel grew throughout the late 19th and early 20th centuries, so an international standard time-zone system was developed, involving a plurality of time standards, i.e. 24 rationally but artificially created time zones around the world. These multiple time zones, which we depend upon so heavily today in our globalised world, do not necessarily align with what is generally referred to as *natural solar-time*. Towns that are relatively close physical neighbours can find themselves located in two adjacent but different time zones, and can thus experience a one-hour differential which seems odd and counter-intuitive from a physio-temporal standpoint. Furthermore, different countries around the world vary in how they implement the international time-zone system, and they may choose different points in the year to switch from standard time to 'summer' or 'daylight saving' time. But a standardised approach to time, universally shared but flexibly applied, remains a useful, even indispensable function in our daily life.

This paper echoes some of the themes associated with the setting of standards for time. It also develops a key theme of the ALTE 2nd International Conference, held in Berlin in May 2005, which explored the challenge of setting quality standards while at the same time sustaining diversity. How we acknowledge and appreciate linguistic diversity, while at the same time setting and maintaining appropriate standards for assessment purposes, remains a key concern for language testers.

What do we mean by 'linguistic diversity'? The phrase embraces the variety or variation we find in language forms and use, variation which is

often shaped by context and purpose. It can include: variation across regional accents and dialects; variation across different professional or personal domains; variation between formal and informal registers; and variation between spoken and written modes of communication. Linguistic diversity also touches upon notions of multilingualism and multiculturalism across groups and within societies; notions of plurilingualism and pluriculturalism at the individual and personal level; and notions of cross-cultural and intercultural understanding within and between nations. The Council of Europe's *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (2001) makes reference to all these dimensions and is committed to sustaining and celebrating linguistic diversity.

However, in an increasingly globalised world we also observe a powerful drive towards the quest for and imposition of national and international standards, and language education is no exception. In the face of this, there is an understandable fear that linguistic diversity and differentiation will be marginalised or overpowered by linguistic *homogenisation*. We sense a tension between, on the one hand, a need for standardisation across variable contexts – to aid transparency, accessibility and communication – and, on the other, a desire to respect, protect and nurture linguistic and cultural diversity so as to appreciate the richness which diversity brings to our human existence and to avoid its loss.

This paper begins by considering the nature of standards and the role that the 'setting of standards' and 'standardisation' play in the business of assessment. It then briefly describes aspects of linguistic variation and language varieties, and considers some implications these have for language teaching, learning and assessment, especially for testing agencies. It touches briefly on the potential factors that shape policy and practice – including factors such as politics and prejudice. The final part of the paper explores how far two frameworks of reference that have recently been proposed within the language testing and assessment community for test validation and evaluation purposes might help us in constructing a principled and pragmatic approach to setting language standards for teaching and assessment.

Standards in testing and assessment

The term 'standard' or 'standards' is a core term within the professional field and discourse of language testing; but it is worth noting that, at least in English, the term embraces a number of different though related concepts. One general English dictionary offers the following definitions:

> **Standard** /stændəd/**, standards.**
> 1. A **standard** is
> 'a level of quality or achievement' . . . 'that is thought to be acceptable'

> 'something used to measure or estimate the quality or degree of something'
> 'a moral principle which affects people's attitude and behaviour'
> 2. **Standard** is also used to describe something which is
> 'usual and normal, rather than being special or extra'
> 'of a normal, basic size, quality or amount'
> 3. **Standard** is 'used to describe spelling, pronunciation, grammar, etc which is generally regarded as correct or acceptable'
> 4. A book described as a **standard** work or text 'is the one most widely read and recommended in a field or the best about a particular subject'
> 5. A **standard** is also 'a flag which is associated with a particular person or group of people'
> (*Collins Cobuild English Language Dictionary* 1987:1,421)

These examples illustrate well how a single word can express a number of different facets of meaning; the insights captured in these definitions are relevant when exploring possible approaches to setting language standards.

Within the narrower, more specialised context of language testing and assessment, the *Dictionary of Language Testing* by Davies, Brown, Elder, Hill, Lumley and McNamara (1999) highlights two, more technical, meanings:

> 1. 'Standard refers to a level of performance required or experienced'
> 2. 'A second sense refers to a set of principles which can be used as a basis for evaluating what language testers do. Standards in this second sense may lead to codification in an agreed set of guidelines or Code of Practice. Such codification indicates a concern to establish professional ethics' (1999:185).

This two-pronged definition highlights the twin demands which confront language educators and measurement specialists:

a) First, the need to *describe the standard* – in terms of the level or quality of performance expected (which can be referred to as *the what*)
b) Secondly, the need to *justify the standard* – in terms of generally accepted principles and professional ethics (which can be referred to as *the why*)

Both these activities – *describing* and *justifying* – are *socially* contextualised. Describing the standard will be largely socially *determined*, drawing on input from applied linguists, language educators, and other language experts and users. Justifying the standard will be socially *oriented*; it will involve constructing and presenting a convincing argument to policymakers and wider society on why a particular standard (or set of standards) is considered appropriate in a given context. Describing and justifying language standards for teaching, learning and assessment requires language testers to think carefully through the issues raised by language diversity and variation. It involves

a sound understanding of language variation, of its nature and extent, and of some specific implications this has for language testers.

The nature and extent of language variation

The nature and extent of language variation and evolution is a well-established and increasingly well-described phenomenon. The global spread of English, for example, over several centuries has led to the emergence of regionally based varieties – British, American, Australian English. More recently, so-called 'new Englishes' have begun to emerge in certain regions, for example, Hong Kong, Singapore and the European Union. The situation for English has been well documented by many writers in the field (see, for example, Brutt-Griffler 2002, Crystal 1995, 1997, Jenkins 2000, 2003, 2007, McArthur 1998, Trudgill and Hannah 1994, and many others). For a helpful and up-to-date overview of this field since the early 1980s, see Bolton (2004).

But English is not the only European language to have experienced such linguistic evolution. Other widely spoken languages in Europe can testify to a similar experience, though perhaps not on such a grand or global scale. French, Spanish and Portuguese all have established or emerging varieties in different parts of the world such as Canada, Mexico, Brazil, or take on a lingua franca role (see, for example, the American Association for Applied Linguistics *Annual Review* on the theme of Lingua Franca Languages, edited by McGroarty 2006). Less widely spoken European languages experience a similar phenomenon. For them, variety may be less wide-ranging geographically but there can still be debate at national and regional level about how closely a particular 'localised' variety of the language does – or does not – align with an accepted or 'acceptable' national standard. Beyond Europe, other languages have assumed or are perceived to be taking on a 'global' role. In March 2008, for example, the Department of Indo-Pacific Languages and Literature at the University of Hawaii, Manoa, hosted a conference on the theme of 'Filipino as a Global Language: Prospects and Future Directions'.

How does linguistic variation manifest itself? At the micro-level, we see it in distinctive phonological, morphological, lexical, syntactic and orthographic features; at a macro-level, variation can be seen in discourse features (to do with rhetorical structure) and in pragmatic features (to do with the socio-cultural context of use).

The function of language variation

The function of language variation is well recognised within applied linguistics: it helps support notions of identity, belonging to a community, or being a member of a particular fellowship. Identity may be regionally based

and reflected in a particular accent or dialect; or it may be more personally or group based, giving rise to what are referred to 'idiolects' or 'sociolects'. Linguistic analysis has also identified variation across high and low language forms (e.g. acrolects, mesolects and basilects), as well as variation according to age, gender or status. In recent years we have learned much about the extent of variation that occurs between language in its *spoken* forms and language in its *written* forms, and the blends of these that appear in new electronic genres, such as emails and blogs.

The relationship between language varieties

How do we explain the relationships between the different regionally based linguistic varieties which exist? Some applied linguists have used biological or geographical analogies, exploiting metaphors such as a tree structure or a wheel (McArthur 1992, Strevens 1980). Most famously, Braj Kachru defined 'world Englishes' as 'the functional and formal variations, divergent sociolinguistic contexts, ranges and varieties of English in creativity, and various types of acculturation in parts of the Western and non-Western world' (Kachru 1992:2). He subdivided varieties of English into three categories or 'circles': *inner*, *outer* and *expanding*, with each circle relating to the notion of norms in a different way. Kachru described some varieties as *norm-providing* (e.g. English as a *native* language in the US, UK, Australia); some as *norm-developing* (e.g. English as a *second* language in India, Nigeria, Malaysia); and others as *norm-dependent* (e.g. English as a *foreign* language in China, Israel, Indonesia). With its focus on 'norms' or standards, Kachru's analysis appears to hold particular relevance for language teachers and testers. However, his model may be too static, failing to reflect the dynamic changes in the nature and status of world Englishes over the past two decades – at least in terms of who or where 'provides' today's norms, and who 'depends' on them? For example, Kachru's model does not easily accommodate the emergence of English as a pan-European lingua franca (Jenkins 2007, Seidlhofer 2001, Seidlhofer, Breiteneder and Pitzl 2006); nor does it account for the development of potential transnational and potentially neutral, non-political varieties such as an international lingua franca English (Crystal 1997, McArthur 1998).

Implications for language teaching and testing

Sophisticated approaches to analysing language, including corpus linguistics studies, have improved our description and understanding of language variation, bringing greater awareness of the issues it raises for language teaching, learning and assessment. For both teacher and tester, there are theoretical and practical issues about *what to teach* – in relation to pedagogy, materials

and training, and *what to test* – in relation to standards, norms, models and judgement criteria (Davies, Hamp-Lyons and Kemp 2003, Elder and Davies 2006, Lowenberg 2000).

As we've learned more about how gender, age, L1 and national identity shape language use, alongside factors of topic, domain, context and purpose, so some of these varietal features have begun to be reflected in language teaching and learning, finding their way into course materials and classrooms, at least for English. Some coursebooks now contain an explicit teaching unit on language varieties and there are accessible scholarly articles and volumes to support teacher training and development (e.g. Jenkins 2003, Kirkpatrick 2007, McKay 2002, Snow, Kamhi-Stein and Brinton 2006). From an assessment perspective, the main practical challenges for language testers relate to *test development* (i.e. selecting appropriate test input) and to *test scoring* (i.e. criteria for evaluating test output, and training of interlocutors and raters). These practical aspects can be located under the superordinate heading of quality and fairness, and I shall come back to this point in more detail later.

Theoretical and practical decisions facing teachers and testers are often complicated by socio-political sensitivities about *whose* language should be the focus of attention. Does one national variety have an inherently higher value than another? If so, which one, and why? And should this variety be imposed as widely as possible for purposes of teaching, learning and assessment? Does such imposition reflect certain political imperatives, or social prejudices? The topic of language varieties is a sensitive one because it touches on issues of community culture and personal identity (see Jenkins' 2007 volume which discusses attitudes and identity in relation to English as a lingua franca); this is true not only for English but also for other languages, particularly in the European context with its constantly shifting socio-political realities.

The information and communications technology (ICT) revolution also has a major impact on our policy and practice. In the 1980s David Crystal (1987) referred to dialects existing on a continuum, with some closer to one another than others. He raised questions about precisely which dialects it is reasonable to expect a language user to cope with: only those dialects which are close to his own, or those at some distance? But nowadays physical or geographical distance can be overcome by communications media such as the worldwide web, and we can experience a form of 'virtual' proximity when seeking information or when communicating with others. Current technologies provide regular exposure to a wide range of language varieties, especially for younger generations of learners; this, combined with the ever-increasing rate of language change in today's fast-moving world, makes decisions about appropriate language models and standards even more complex for teachers and testers.

What do we learn from a survey of current practice among test providers?

In 2000 Cambridge ESOL decided to review and reflect on its policy and practice on the varieties of English used in its tests (see Taylor 2006 for some discussion of the issues). The review included a survey of documentation from other international providers of English language proficiency tests to see what they said about their stimulus materials, test task design, assessment criteria, standard/norms, and rater training. The survey indicated a striking diversity of approach across different English language testing agencies, with few test providers seeming to offer a clear rationale for their policy and practice (Taylor 2001, 2002).

In 2004 the survey was broadened to include test providers of languages other than English. The ALTE partnership offered a unique context for conducting a small-scale study of perceptions, policy and practice among a selection of European language testing agencies; findings from this study were reported at the 2005 ALTE Conference in Berlin and published in a proceedings volume (Taylor 2008). Across a sample of 10 European test providers representing nine different European languages (Basque, Czech, Dutch, English, French, German, Italian, Portuguese [x2] and Welsh), it was clear that there were perceptions and experience in common; however, there was also evidence of some diversity in attitude and approach. Organisational policy and practice were shaped by factors such as the context and purpose for a given test (i.e. where the test is being used, what/who it is designed for, etc.). Other considerations such as validity, reliability, impact and practicality clearly played their part too, along with the notion of 'test usefulness' or 'fitness for purpose'. Although the words 'equal access' and 'fairness' did not appear explicitly in the survey responses, there was evidence that these too were considerations. At a more macro-level, it was clear that historical tradition and socio-political factors also influenced choice of language standards.

Language testing professional standards and codes of practice

The last 15–20 years have seen the publication of a number of professional standards and codes of practice for the language testing profession, so it is reasonable to ask what help these offer test producers when developing policy and practice on language diversity.

One of the earliest codes, originally published in 1988 and updated in 2004, is the US *Code of Fair Testing Practices in Education*. The Code is aspirational rather than mandatory, presenting standards for educational test developers and users in four key areas: developing and selecting appropriate tests; administering and scoring tests; reporting and interpreting test results; and inform-

ing test takers. It offers no advice on language standards; in fairness, however, the Code is concerned with testing practice in general, not language testing in particular. This US Code was the inspiration for the ALTE Code of Practice drawn up in 1994, which sought to make explicit the standards the ALTE members aim to meet; but even here there is no specific mention of language standards, nor any guidance on how to address the issue of language diversity.

The 1999 *Standards for Educational and Psychological Testing* produced by AERA/APA/NCME contain a chapter on 'Fairness in testing and test use' addressing four dimensions: absence of bias; equitable treatment of all examinees; equality of testing outcomes; equity in opportunity to learn the material covered. Here again, there is no explicit reference to the issue of linguistic variety except in relation to the background languages (i.e. L1) of test takers. The ETS *Standards for Quality and Fairness*, published in 2002, understandably focus on general rather than language assessment matters; however, the ETS *Fairness Review Guidelines* (2003) do contain a few references to linguistic diversity issues, advising against the use of 'regionalisms', or stereotypes associated with dialect or language usage; and the ETS *International Principles for Fairness Review of Assessments* (2004) advise on the use of American or British English according to the country for which the test is intended. EALTA's *Guidelines for Good Practice in Language Testing and Assessment* (2006) are intended for training teachers in testing and assessment, for those involved in classroom assessment or in the development of tests in national or institutional testing units or centres. Like the ETS and ALTE Standards, the *Guidelines* stress a number of general principles, including: respect for the students/examinees, responsibility, fairness, reliability, validity, and collaboration among the parties involved. Once again, they offer no explicit guidance on the specific issue of language standards. Similarly, the *ILTA Draft Code of Practice* (Version 3, 2005) refers to basic considerations for good testing practice but does not address language standards specifically.

Language test developers regularly find themselves having to set and maintain language standards for assessment purposes and at the same time acknowledge and reflect the reality of linguistic diversity. If the ethical codes and good practice guidelines set out by the professional community do not help them to balance and reconcile these twin demands, where can test developers go for guidance on their current policy and practice? And who will help inform its future development as the languages taught and tested continue to evolve, and as new or established language varieties assume an increased social, political or educational significance?

Some useful frameworks of reference

The remainder of this paper explores two frameworks of reference that have emerged recently within the profession which may help language test

developers to think critically and creatively about the setting of language standards with regard to language diversity. The two frameworks adopt differing but complementary perspectives: the first takes a primarily *ethical* perspective, with a focus on ensuring and demonstrating equity and fairness; the second takes a more *technical* perspective, focusing on the collection of evidence for test validation purposes.

The Test Fairness and Test Context Frameworks (Kunnan 2004, 2008)

Antony Kunnan is one of the writers within the field who has sought to address the issue of language standards and diversity explicitly within the broader context of fairness. In his opening paper 'Fairness and justice for all' in Volume 9 of the *Studies in Language Testing* series, he writes about fairness at the test writing stage: 'In addition, decisions regarding the language standard(s) (or dialects) that are to be adopted for the test need to be made by the developers and writers' (2000:8). Drawing upon the chapter in the 1999 *Standards* on 'Fairness in testing and test use', Kunnan conceptualises a *Test Fairness Framework (TFF)* which views fairness in terms of the whole system of a testing practice rather than just a test in isolation. He uses this Framework to consider five test quality categories of: *validity*, *absence of bias*, *access*, *administration* and *social consequences* (2004:37–39).

Language variety issues can be systematically evaluated against these five categories. For example, he explicitly mentions consideration of 'language dialect' under the test qualities of *content and construct validity*, and 'choice of dialect' is linked to striving for *absence of bias*. Though not explicitly mentioned by Kunnan, the *educational/personal access – opportunity to learn* category challenges us to consider learners' or test takers' likely familiarity with particular accents and dialects, or with other types of variation such as formal/informal registers or spoken/written forms. Similarly, diversity issues relating to interlocutors and raters for speaking and writing assessment could be evaluated in relation to *uniformity or consistency in administration*. The fifth test quality category is *social consequences* and this too opens up a space for considering the potential washback and impact of including – or not including – language dialects or other types of variation in a test. Thus, the TFF can help language testers to consider, in a systematic manner, relevant theoretical and practical matters relating to language diversity to determine how these are addressed in a test.

More recently, Kunnan has developed an additional tool for evaluating tests and testing practice (Kunnan 2008). His *Test Context Framework (TCF)* is a complementary framework to the TFF and it prompts us to examine tests and testing practice from a wider perspective to determine whether and how these tests are beneficial or detrimental to society. The TCF refers to the

collection of traditions, histories, customs, and academic and professional practices, and social and political institutions of a community, i.e. the political and economic, the educational, social and cultural, the technological and infrastructure, and the legal and ethical contexts of a community in which a test operates. All of these contexts have potential implications for decisions about setting language standards and their relationship to language diversity. The TCF can be used to consider critically the political and social assumptions about language diversity that may shape thinking and practice among language teachers and testers, including traditional assumptions or historical prejudices that may need challenging and rethinking.

The Socio-Cognitive Framework (Weir 2005, Shaw and Weir 2007)

An alternative approach to considering the issues can be found in work by Cyril Weir. His Framework for evidence-based test validation adopts a more technical, validity-oriented perspective than the Kunnan TFF and TCF frameworks. It's important to note, however, that although Kunnan and Weir start from slightly different standpoints, both their approaches are concerned with collecting appropriate validation evidence about a given test's qualities to support claims about its usefulness and fairness. Weir proposes six conceptual categories as follows to help organise our thinking: *test taker*; *cognitive validity*; *context validity*; *scoring validity*; *consequential validity*; and *criterion-related validity*. Once again, these provide us with a useful heuristic for reflection and action in the area of language diversity.

Under *test taker* we are prompted to consider the physical, psychological and experiential characteristics of the learner or test-taker population of interest, i.e. their age, language background, learning experience, and what this implies for the language standards we set. If our test takers' language learning experience has been limited to written forms only, should the test include a speaking component?

Under *cognitive validity* we can consider the mental processes required to complete the test tasks and whether they are interactionally authentic. Would it be appropriate, for example, to expect young language learners to cope with tasks involving awareness of how language varies from formal to informal registers, or is this inappropriate given their level of developmental maturity?

Under *context validity* we can consider the contextual features of a test task and its administration. Would it be reasonable, for example, to expect test takers to cope with different speaker accents in a listening test, or should accents be restricted to those to which they have been most exposed in the past, or to which they will be exposed in the future target language use (TLU) context? One respondent in the ALTE 2005 survey (Taylor 2008) speculated

that the use of increasing variation in local media alongside 'standard' forms in the national media (newspapers, TV, etc.) meant the population as a whole was becoming more accustomed to dialects, with communication across the dialect continuum becoming easier as a result. This suggests that policy and practice on test content and the contextual features of test tasks may need to change over time as the surrounding linguistic landscape itself evolves. Test content sampling will ideally be as representative as it can be of the TLU domain. This may have implications for the level of variation that can reasonably be considered acceptable or desirable across different modes (spoken and written), or registers (informal and formal), as well as other types of variation such as accentedness. The notion of the 'main host language of communication' may be a helpful one to consider here. For example, in the case of IELTS (International English Language Testing System), an international English proficiency test used to assess the language level needed for study, work or training in English speaking environments, the test tasks are prepared by an international team of test writers drawn from the TLU context (i.e. UK, Australia, New Zealand). Consequently, test input, especially in the listening test, reflects features of the English varieties used in the TLU domain. The design specification for any test will ideally draw on some form of linguistic description of the target language (including extent of variation). This may be easier in situations where the TLU context is relatively easily defined. It can prove more difficult, however, where the TLU context is far broader and lingua-culturally heterogeneous, or is perhaps still awaiting description or codification, e.g. the use of English as a lingua franca between non-native English speakers.

Under *scoring validity* we consider how far we can depend on the scores of the test. For example, might reliable assessment of performance on a writing test be threatened by rater uncertainty over whether British, American or some other English is acceptable? How should raters treat variable features of candidates' written or spoken production (e.g. spelling or pronunciation) which may reflect a variety they have learned or grown up with? Once again, clarifying the test construct and purpose can help us here. Test developers need to be clear about the focus of their assessment and the degree of precision they require. We may be prepared to accept differential standards for different modes of communication; e.g. greater flexibility in evaluating candidates' spoken language (where variation tends to be the norm) and more stringent requirements in written production (where conformity to a standard is more likely or more desirable). At the very least, marking criteria (e.g. in relation to spelling requirements) must be as transparent as possible not just for raters, but also for test candidates; test takers are often aware of differences across language varieties and sometimes fear they will be penalised for using the 'wrong' lexical item or spelling convention.

Under *consequential validity* we can consider what impact the test has on its various stakeholders. For example, we might ask what value messages the inclusion of only so-called 'native speaker' varieties sends out to the wider world. How will this impact on teaching and learning? And finally, under *criterion-related validity* we can consider external evidence of the test's meaningfulness and usefulness. It may be important to think carefully about setting a native-speaker standard as a goal in a world where plurilingualism is increasingly the norm. This touches upon the controversial question of whether NS or NNS examiners are better qualified to evaluate proficiency. The reality must surely be that all interlocutors and raters – both NS and NNS – need to be suitably qualified, and to receive initial training and ongoing standardisation for their work. Lowenberg (2000) has suggested that some awareness of potential divergence in norms across so-called native/non-native varieties should be an essential part of any rater's expertise, whether or not they are a 'native speaker'.

Conclusion

This paper suggests that language testers need to develop a principled and well thought out approach to their policy and practice on setting language standards and accommodating linguistic diversity, rather than allow the approach to be shaped primarily by tradition, politics or prejudice. Tradition and politics, of course, cannot be ignored and they will need to be considered within the decision-making process, as indicated in Kunnan's Test Context Framework; but prejudice – whether it is national or personal, whether it is conscious or unconscious – is likely to be a poor foundation for sound policy and defensible practice. When internal and external test stakeholders ask the question *How do you deal with linguistic diversity when you set the language standards for your tests?,* test developers need to be able to respond with clarity, confidence and conviction for various reasons.

One reason is that perceptions about the 'ownership' of a given language can evolve over time, because language is so closely linked with issues of socio-cultural identity, culture and power. We only need to look at the changed and changing face of Europe in our own time to know that this is true. Critical linguists warn of the dangers of 'linguistic imperialism' or 'linguicism' (Pennycook 1994, Phillipson 1992), of allowing one strong language variety to dominate and marginalise other weaker varieties. Some perceive the control or ownership of the English language to be transferring away from the traditional white, Anglo-Saxon, Protestant L1 speaker communities of Britain, the USA, Australia, etc., and this undoubtedly has implications for future teaching and testing.

A second reason is the rapid rate of language change today (Aitchison 2001, Crystal 2000), perhaps because of the mass movement of language

users as a result of global business, political instability, economic or environmental deprivation, or international tourism. Furthermore, language users are increasingly engaged in a form of 'virtual migration' made possible by modern information and communications technology, and this potentially reshapes the linguistic landscape.

Thirdly, in the context of language education more specifically, we are seeing steady growth in the 'localisation' of language teaching and assessment: locally published syllabuses, curricula, and teaching materials, as well as the development of locally generated standards for teaching, learning and assessment. There is also a growing focus on the teaching of oral communication skills (listening and speaking) and linguistic flexibility and variation are perhaps most manifest in the oral mode. The trend towards 'localisation' sits in tension with the increasing 'globalisation' of educational and employment opportunities for which transparent and accessible international standards – both written and spoken – are sought. There needs to be a way of counterbalancing these twin trends.

Fourthly, today more than ever before, there exist sophisticated tools to analyse and describe the nature of linguistic variation, for example through corpus-based studies of spoken and written language. Such advances make possible the study and codification of less widely spoken languages and linguistic varieties as well as just the 'big' languages. Findings from specialised corpora of language varieties (spoken/written, child/adult, NS/NNS) are starting to feed into approaches to language teaching and language testing. Testing agencies have a valuable role to play in building corpora of written and spoken language test performances which can be analysed for insights into linguistic diversity (see Taylor and Barker 2008 for more on this).

Fifthly, sound policy and practice on linguistic diversity matter for language testers because of current concerns with validity; language testers must pay due attention to the quality standards they lay claim to. The strong focus nowadays on accountability and fairness impacts on professional and public attitudes to tests and test use; this was directly reflected in the theme of the ALTE Cambridge conference.

Language testing organisations are sometimes criticised for taking insufficient account of linguistic diversity in their testing practice, or for failing to take more of a lead in promoting recognition of language varieties (Jenkins 2006, Lowenberg 2000). Such criticism is understandable given the high-stakes role that language testing plays in so many parts of the world but it does not always readily acknowledge the complexities involved in dealing with language diversity and linguistic variation in the assessment context. There are particular challenges when dealing with a large and/or highly heterogeneous test population, e.g. an international test candidature, or a population of test takers with a potential age range from 17 to 70. Not surprisingly, perhaps, the debate in recent years on language varieties and assessment has been a

vigorous and stimulating one among applied linguists and language testers. But too often it seems to seek simple answers to complex questions, expecting absolute truths and clear guidelines amidst shifting social realities and pragmatic compromises. It is too easy for the debate about language standards and linguistic diversity to become trapped in endless and unproductive discussions: discussions about political power and national ownership, about the relative merits of the native speaker versus non-native speaker model, about the evils of linguistic imperialism and the need for linguistic democracy, or about the importance of maintaining traditional, even absolute language standards, for reasons of 'correctness' or 'prestige'. The problem with this sort of discourse is that ultimately it fails to provide any principled and workable solutions to the practical real-world challenges faced by language testers and others involved in language education.

This paper seeks to move us beyond the limitations of the traditional socio-political debate. Instead, it proposes a framework of reference for contextualising the issues so that individual language testers and language testing organisations can work out the policy and the practice for themselves in a pragmatic but principled way, a way that is consistent with current ethical and professional understanding in the field. Antony Kunnan's Test Fairness and Test Context Frameworks offer useful tools for considering how to set language standards in the context of test evaluation; his approach helps language testers to assemble and articulate the necessary evidence in support of a test's utilisation argument. Similarly, the socio-cognitive framework presented by Cyril Weir offers a mechanism for highlighting where language diversity issues arise in testing practice and how these might be addressed in a transparent and principled way to construct a validity argument. Both these frameworks build directly upon a unified view of validity, reconceptualised during the 1990s in the work of Messick (1989, 1996) who advanced a critical role for the value implications and social consequences of language testing and assessment. Indeed, it is unlikely the ALTE conference in Cambridge would have taken place at all were it not for Messick and others working in a similar vein over the past two decades.

Being able to explain the approach to setting language standards in accordance with a principled framework of reference such as those described here should allow language test developers to refute accusations that their policy and practice are determined by politics or prejudice, whether these are personal, group, institutional or national in nature. It should also aid the urgent task of improving 'assessment literacy' among the many testing stakeholders who exist today. A better public understanding of language testing principles and practice, including their benefits and limitations, will surely enhance the positive social and educational impact of assessment.

This paper has sought to explore some of the issues and challenges that language testing agencies face in setting language standards for teaching and

assessment, and to suggest accessible and effective approaches to working through the issues. By way of conclusion, it seems appropriate to return to my introduction to this paper where I commented on the development of a standardised approach to time in public life nationally and internationally.

Sociologists and social historians have noted that human beings have the capacity to describe points in time in different, often highly creative ways. Here is one such example from a fictional work: 'When I was a younger man – two wives ago, 250000 cigarettes ago, 3000 quarts of booze ago. . .' (Vonnegut 1963:11, cited in Zerubavel 1982). This manner of talking about time, what Schegloff (1972:116) refers to as a 'temporal formulation', is rich in meaning at the personal level for the writer, and it may well carry figurative meaning for the reader. But such an approach is likely to be of little use at the collective level where we need to be able to draw a distinction between psychological and sociological perspectives on temporal reference or time-reckoning. Emile Durkheim, one of the founding fathers of sociology, reminds us that 'What the category of time expresses is a time common to the group, a *social* time, so to speak' (1965:23, cited in Zerubavel 1982). The process to standardise time in the 19th century was just such a social endeavour, undertaken for the benefit of the collective. As Stephens astutely observes, 'It is a truism that modern life runs by the clock. Clocks synchronize our communal activities, and that they do so is at once useful and tyrannical' (1989:1).

Perhaps we can draw a parallel here between the socially constructed nature of standard time as we know and experience it, and the socially constructed nature of language standards for teaching and assessment. This is not a new idea: Davies (1991, 2003) has written extensively on the 'native speaker concept' in the field of applied linguistics and language testing, and he refers to the notions of native speaker and standard language in terms of their social construction and functionality. In effect, perhaps all standards used in human society, including language standards, are *socially constructed* in some sense or other; very few are God-ordained – some might say only two: 'Love God' and 'Love your neighbour'. Language standards, like most other standards, are socially and culturally determined, designed to serve the community where they are developed and to meet a specific set of needs.

In his fascinating paper on the standardisation of time (one of several papers drawn on for this paper, including Bartky 1989 and Stephens 1989), the American sociologist Eviatar Zerubavel views the development of standard time as arising from the socio-historical need to synchronise different communities and countries with one another. The outcome, he claims, was a time-zone system which helped to solidify 'organic' ties among people, a system which he believed manifested the twin virtues of 'interdependence and complementary differentiation' (Zerubavel 1982:21).

In setting language standards for teaching and assessment, perhaps we

too should be striving to achieve a balance between the twin virtues of inter-dependence and complementary differentiation.

References and further reading

Aitchison, J (2001) *Language Change: Progress or decay?* (3rd ed) Cambridge: Cambridge University Press.

AREA/APA/NCME (1999) *Standards for Educational and Psychological Testing*, Washington, DC: Author.

Association of Language Testers in Europe (1994) *ALTE Code of Practice.*

Bartky, I R (1989) The adoption of standard time, *Technology and Culture*, 30 (1), University of Chicago Press, 25–56.

Bolton, K (2004) World Englishes, in Davies, A and Elder, C (Eds) *The Handbook of Applied Linguistics*, Oxford: Blackwell, 367–396.

Brutt-Griffler, J (2002) *World English*, Clevedon: Multilingual Matters Ltd.

Code of Fair Testing Practices in Education (2004) Washington, DC: Joint Committee on Testing Practices.

Collins (1987) *Collins Cobuild English Language Dictionary*.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Crystal, D (1987, 1995) *The Cambridge Encyclopedia of the English Language*, Cambridge: Cambridge University Press.

Crystal, D (1997) *English as a global language*, Cambridge: Cambridge University Press.

Crystal, D (2000) *Language Death*, Cambridge: Cambridge University Press.

Davies, A (1991) *The native speaker in applied linguistics*, Edinburgh: Edinburgh University Press.

Davies, A (2003) *The native speaker – myth or reality?* Clevedon: Multilingual Matters Ltd.

Davies, A, Brown, A, Elder, C, Hill, K, Lumley, T and McNamara,T (1999) *Dictionary of Language Testing*, Cambridge: UCLES/Cambridge University Press.

Davies, A, Hamp-Lyons, L and Kemp, C (2003) Whose norms? International proficiency tests in English, *World Englishes*, 22, Oxford: Blackwell Publishing Ltd, 571–584.

Durkheim, E (1965) *The Elementary Forms of the Religious Life*, New York: Free Press.

Educational Testing Service (2002) *ETS Standards for Quality and Fairness*, Princeton, NJ: ETS.

Educational Testing Service (2003) *ETS Fairness Review Guidelines*, Princeton, NJ: ETS.

Educational Testing Service (2004) *ETS International Principles for Fairness Review of Assessments*, Princeton, NJ: ETS.

Elder, C and Davies, A (2006) Assessing English as a Lingua Franca, in McGroarty, M (Ed.) *Annual Review of Applied Linguistics – An official journal of the American Association of Applied Linguistics*, Cambridge: Cambridge University Press, 282–301.

European Association of Language Testing and Assessment (2006) *EALTA Guidelines for Good Practice in Language Testing and Assessment*.

International Language Testing Association (2005) *ILTA Draft Code of Practice*, Version 3.
Jenkins, J (2000) *The Phonology of English as an International Language*, Oxford: Oxford University Press.
Jenkins, J (2003) *World Englishes: a resource book for students*, Routledge.
Jenkins, J (2006) The spread of EIL: a testing time for testers, *ELT Journal*, 60 (1), Oxford: Oxford University Press, 42–50.
Jenkins, J (2007) *English as a Lingua Franca: Attitude and Identity*, Oxford: Oxford University Press.
Kachru, B B (1992) World Englishes: approaches, issues, resources, *Language Teaching*, 25, 1–14.
Kirkpatrick, A (2007) *World Englishes: implications for international communication and English language teaching*, Cambridge: Cambridge University Press.
Kunnan, A J (2000) Fairness and justice for all, in Kunnan, A J (Ed.) *Fairness and validation in language assessment*, Cambridge: UCLES/Cambridge University Press, 1–13.
Kunnan, A J (2004) Test fairness, in Milanovic, M and Weir, C (Eds) *European Language Testing in a Global Context,* Cambridge: UCLES/Cambridge University Press, 27–48.
Kunnan, A J (2008) Towards a model of test evaluation: using the Test Fairness and Test Context Frameworks, in Taylor, L and Weir, C J (Eds) *Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference May 2005*, Cambridge: UCLES/Cambridge University Press, 229–251.
Lowenberg, P H (2000) Non-native varieties and issues of fairness in testing English as a world language, in Kunnan, A J (Ed.) *Fairness and validation in language assessment,* Cambridge: UCLES/Cambridge University Press, 43–59.
McArthur, T (1992) Models of English, *English Today* 32.
McArthur, T (1998) *The English Languages*, Cambridge: Cambridge University Press.
McGroarty, M (2006) (Ed) *Annual Review of Applied Linguistics – An official journal of the American Association of Applied Linguistics*, Cambridge: Cambridge University Press.
McKay, S L (2002) *Teaching English as an International Language*, Oxford: Oxford University Press.
Messick, S (1989) Validity, in Linn, R (Ed.) *Educational Measurement*, New York: Macmillan, 13–103.
Messick, S (1996) Validity and washback in language testing, *Language Testing* 13 (3), 241–256.
Pennycook, A (1994) *The cultural politics of English as an international language,* London: Longman.
Phillipson, R (1992) *Linguistic imperialism*, Oxford: Oxford University Press.
Schegloff, E A (1972) Notes on a conversational practice: formulating place, in Sudnow, D (Ed) *Studies in Social Interaction*, New York: Free Press, 75–119.
Seidlhofer, B (2001) Closing a conceptual gap: the case for a description of English as a lingua franca, *International Journal of Applied Linguistics*, 11 (2), 133–58.
Seidlhofer, B, Breiteneder, A and Pitzl M L (2006) English as a lingua franca in Europe: Challenges for applied linguistics, in McGroarty, M (Ed.) *Annual*

Review of Applied Linguistics – An official journal of the American Association of Applied Linguistics, Cambridge: Cambridge University Press, 3–34.

Shaw, S D and Weir, C J (2007) *Examining Writing: Research and practice in assessing second language writing*, Cambridge: UCLES/Cambridge University Press.

Snow, M A, Kamhi-Stein, L D and Brinton, D M (2006) Teacher training for English as a lingua franca, in McGroarty, M (Ed.) *Annual Review of Applied Linguistics – An official journal of the American Association of Applied Linguistics*, Cambridge: Cambridge University Press, 261–281.

Stephens, C (1989) 'The most reliable time': William Bond, the New England railroads and time-awareness in 19th century America, *Technology and Culture*, 30 (1), University of Chicago Press, 1–24.

Strevens, P (1980) *Teaching English as an International Language*, Oxford: Pergamon.

Taylor, L (2001) *Assessing learners' English: but whose/which English(es)?*, paper presented at the UK Language Testing Forum in Nottingham, November 2001.

Taylor, L (2002) Assessing learners' English: but whose/which English(es)? *Research Notes* 10, University of Cambridge ESOL Examinations, 18–20.

Taylor, L (2006) The changing landscape of English: implications for language assessment, *ELT Journal*, 60 (1), Oxford: Oxford University Press, 51–60.

Taylor, L (2008) Language varieties and their implications for testing and assessment, in Taylor, L and Weir, C J (Eds) *Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference May 2005*, Cambridge: UCLES/Cambridge University Press, 276–295.

Taylor, L and Barker, F (2008) Using Corpora in Language Assessment, in Shohamy, E and Hornberger, N (Eds) *Language Testing and Assessment, Volume 7, Encyclopedia of Language and Education*, New York: Springer Science+Business Media LLC, 241–254.

Trudgill, P and Hannah, J (1994) *International English: a guide to varieties of standard English* (3rd edn), London: Edward Arnold.

Vonnegut, K (1963) *Cat's Cradle*, New York: Dell.

Weir, C J (2005) *Language Testing and Validation: an evidence-based approach*, Basingstoke: Palgrave Macmillan.

Zerubavel, E (1982) The standardization of time: a sociohistorical perspective, *American Journal of Sociology*, 88 (1), 1–23.

9 Using learner language from corpora to profile levels of proficiency: insights from the English Profile Programme

John A Hawkins
University of Cambridge, RCEAL and University of California, Davis
Paula Buttery
University of Cambridge, RCEAL

Abstract

Several decades of practical work on language testing and teaching have led to the six proficiency levels of the Common European Framework of Reference for Languages (CEFR). In this paper we ask the question: how much of the grammar, lexicon and usage conventions of English do learners actually know at each of these levels? The work we report on is based on an empirical examination of the Cambridge Learner Corpus. The accessibility of items has been enhanced through part-of-speech tagging and parsing, permitting searches to be conducted that go beyond individual words. Our ultimate goal is to identify 'criterial features' that distinguish the different proficiency levels from one another, as well as first language transfer effects. The English Profile Programme differs from earlier 'profiling' studies by being empirically based in this way and by controlling for different first languages. Examples are given of findings to date, and some practical benefits of this work are outlined.

Introduction

Work on language testing and teaching over many years has led to the six proficiency levels of the Common European Framework of Reference (CEFR), as summarised in the Council of Europe's 2001 document *Common European Framework of Reference for Languages: Learning, teaching, assessment* (Cambridge University Press). These levels are given in (1):

(1) The CEFR levels:
A1 Breakthrough
A2 Waystage
B1 Threshold
B2 Vantage
C1 Effective Operational Proficiency
C2 Mastery

In this paper we ask the question: how much of the grammar, lexicon and usage conventions of English do learners actually know at each of these levels? Attempts to describe the defining characteristics of each of the levels hitherto have been rather general or have been couched in functional terms (the 'Can Do' statements). Greater precision can be achieved through the use of electronic corpora. The collaborative work we report on here is based on an empirical examination of the Cambridge Learner Corpus. At the time of going to press the ever-expanding Cambridge Learner Corpus consists of over 30 million words of text. The corpus contains candidates' responses to Cambridge ESOL examination written papers, which require candidates to produce an extended piece of prose. The corpus includes three types of examination that, taken together, cover the CEFR levels from A2 to C2: the Main Suite (Certificate of Proficiency in English, Certificate of Advanced English, First Certificate in English, Preliminary English Test, Key English Test); the Business Suite (Business English Certificate Higher, Business English Certificate Vantage, Business English Certificate Preliminary); and the International English Language Testing System (IELTS). Each examination script has been transcribed by Cambridge University Press and approximately half of the corpus error-coded. The scripts have been anonymised but key meta-data for each candidate (such as age, gender and first language) have been retained. Subsequently, the accessibility of linguistic items within the corpus has been enhanced through part-of-speech tagging and parsing, permitting searches to be conducted that go beyond individual words. New codes have been entered into the data that facilitate these searches and that enable us to look for distinguishing features of each proficiency level and for first language transfer effects.

If we can give a good description of the learners' linguistic abilities as their learning progresses, we can contribute to the major goal of the English Profile Programme (EPP) initiated by the Cambridge ESOL division of Cambridge Assessment, which is to provide Reference Level Descriptions for English for all six CEFR levels. Specifically, those of us who are working on this programme at the Research Centre for English and Applied Linguistics (RCEAL) in Cambridge aim to deliver two products to the EPP:

- a set of 'criterial' features that characterise and distinguish the six CEFR levels with respect to English, and

- an assessment of the impact of different first languages (L1s) on performance at each of the levels, and of their interaction with the criterial features.

The 'Can Do' statements describe the functions that learners can use language for at the different levels. For example, we are told that learners at B1 can 'express opinions on abstract/cultural matters in a limited way or offer advice within a known area'. Learners at B2 'can follow or give a talk on a familiar topic'. And learners at C1 'can contribute effectively to meetings and seminars within own area of work or keep up a casual conversation with a good deal of fluency'.

But learners who perform each of these tasks may be using a wide variety of grammatical constructions and words in order to do so, and the ability to 'do' the task does not tell us how exactly the learner does it and with what grammatical and lexical properties of English (or of other target languages). We need to know, for each of the levels:

- which grammatical constructions are used?
- which words? and
- which syntactic and morpho-syntactic rules are applied and with what level of success?

The reason this is important is because knowing a language and being a native speaker means that you have acquired thousands and thousands of properties of English or Spanish or Japanese, etc., including the following:

- the sounds of the language
- meaningful units or morphemes
- words (e.g. the nouns and the verbs)
- basic grammatical constructions
- productive syntactic and morpho-syntactic rules
- exceptions to some of these, i.e. lexical idiosyncrasies
- and so on.

As learners progress, they master more and more of these properties, and move closer to the knowledge of a native speaker. What is fascinating, scientifically, about the practical task of examining learner English is that examiners appear to have built up clear intuitions over several years of experience about B1 and B2 and C1 levels of English, etc., and about the properties of English that learners know and use at each of these levels. And evidence for this comes from the fact that examiners generally show high levels of agreement among themselves in making their practical assessments.

The EPP is an attempt to describe these properties of learner English at each level, and to build on earlier research in this direction by John Trim, the founding head of Linguistics at Cambridge University, and by his collaborators in the Council of Europe.

Background to the EPP

The EPP has been based, initially at least, on an examination of the Cambridge Learner Corpus (CLC). The CLC currently comprises over 30 million words of written learner data, roughly half of which is coded for errors. The CLC was originally searchable only lexically, i.e. on the basis of individual words. But within the context of the EPP the search capability has been expanded and the CLC has been tagged for parts of speech and parsed using the Robust Accurate Statistical Parser (RASP) (Briscoe, Carroll and Watson 2006). This is an automatic parsing system incorporating both grammatical information and statistical patterns, and details of its operation are summarised very briefly in this section and in more detail below.

The CLC's error codes classify over 70 error types. A small sample is given in (2), together with sentences exemplifying each:

(2) Sample Error Codes in the CLC

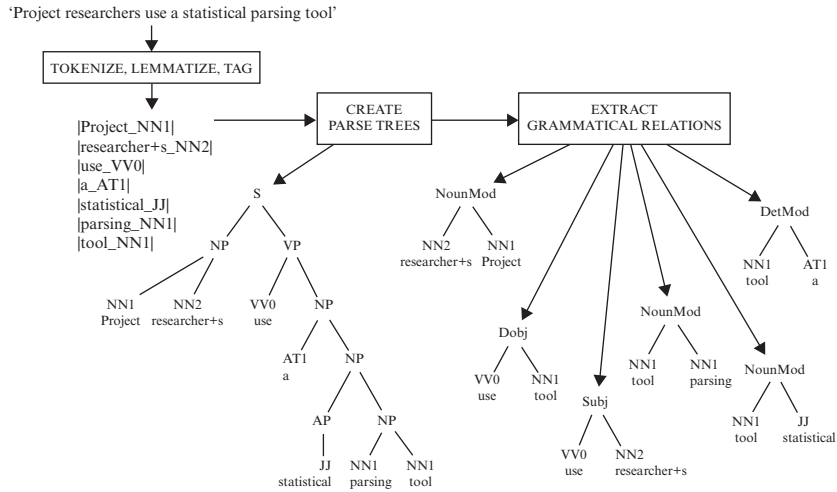
RN	**Replace noun**	Have a good travel (journey)
RV	**Replace verb**	I existed last weekend in London (spent)
MD	**Missing determiner**	I spoke to President (the)
		I have car (a)
AGN	**Noun agreement error**	One of my friend (friends)
AGV	**Verb agreement error**	The three birds is singing (are)

The CLC also contains data from numerous typologically and genetically different first languages.

When the RASP system is run on raw text, such as the written sentences of the CLC, it first marks sentence boundaries and performs a basic 'tokenisation'. Part-of-speech tags are assigned on a probabilistic basis. The text is then 'lemmatised', based on the tags assigned to word tokens. For each sentence a parse forest representation is generated containing all possible parse trees and subanalyses, with their associated probabilities. And a weighted set of grammatical relations is extracted associated with each parse tree. These operations are shown in Figure 1 for the sample sentence 'Project researchers use a statistical parsing tool', using just one illustrative parse tree and its associated grammatical relations.

The second author of the paper, Paula Buttery, is a specialist in computational linguistics and she is the one who is primarily responsible for the computational aspects of the project that are conducted at RCEAL. The RCEAL team also uses research on complexity metrics in order to better understand the developmental stages in learning, as summarised and discussed in Hawkins (1994, 2004, 2009). Learner languages are 'simpler' at first, becoming more 'complex' later. Formal metrics enable us to give some precision and content to this observation. RCEAL also has specialists in first and

Figure 1 'Project researchers use a statistical parsing tool'



second language acquisition, such as Henriette Hendriks and Teresa Parodi. One question they are investigating is the role of the first language in second language acquisition, namely: How do the different first languages of English learners impact their learning of English? What exactly is transferred from L1 to L2? And how does L1 material combine with L2 in the interlanguage?

To properly assess questions of transfer, we need background information about language typology and language universals. How exactly do Japanese and Chinese and Spanish and English differ from each other as language types, grammatically and lexically? And how do structural differences between Chinese and Japanese and Spanish impact the amount and the nature of what is transferred into the Chinese–English interlanguage, the Spanish–English interlanguage, and so on? RCEAL has experts in diverse languages such as Chinese (Henriette Hendriks), Spanish (Teresa Parodi), and in typology and universals (Hawkins).

Overall, the research we describe in this paper is interdisciplinary and is attempting to contribute to the Common European Framework and its assessment levels by combining:

- computational work on the CLC
- psycholinguistic complexity metrics in addition to grammatical and lexical analysis
- acquisition theory and work on language transfer, and
- typological work on cross-linguistic comparison.

One of the strengths of an empirically based approach is that we can now focus not just on errors (i.e. on what learners get WRONG), but on what they

get RIGHT. Using the corpus we can quantify, for each learning stage, how many of the thousands of properties that constitute knowledge of English learners actually use. And we can measure how their linguistic performance gradually improves relative to that of native English speakers. In order to compare the learner data with actual English usage by native speakers we use the British National Corpus (BNC). The BNC comprises 100 million words of modern British English, from a wide range of sources and text types (90 million written, 10 million spoken). The BNC has now been tagged and parsed using the same automatic parsing system (RASP) that we applied to the CLC, which makes exact comparison between them possible.

The EPP is, of course, a project that focuses on English as a second language. But if this research programme can be successfully implemented for English, then we will have a method and a tool that can be applied to all the other languages of Europe, making it a project of far wider potential significance.

Some general patterns and principles of second language acquisition

A number of general patterns and principles are now emerging from this corpus study, some of which will be summarised and illustrated here. Identifying these patterns and principles in the data, and testing certain initial hypotheses that were defined at the outset of this project (see www.english-profile.org/documents/UCLES_RCEAL_Projects.pdf) is an important first stage in identifying criterial features and transfer effects at each level. Three principles will be discussed here: Frequency vs. Infrequency, Structural Simplicity vs. Structural Complexity, and Maximise Transfer.

Frequency vs. Infrequency

This principle is summarised in (3):

(3) Frequency (F) vs. Infrequency (I)

More frequent properties in L2 are more easily acquired, in general: fewer errors, more of the relevant L2 properties learned, and earlier acquisition.

Infrequency (I) has the reverse effects: more errors, fewer of the relevant L2 properties learned, later acquisition.

In other words, we expect to see, and we do see, a disproportionate use of frequent items and properties in early L2 English, moving gradually to more native-like L1 English frequency/infrequency balances.

Consider just one illustration at this point. Learning English nouns and verbs with high frequencies of use should be easier, in general, than learning those

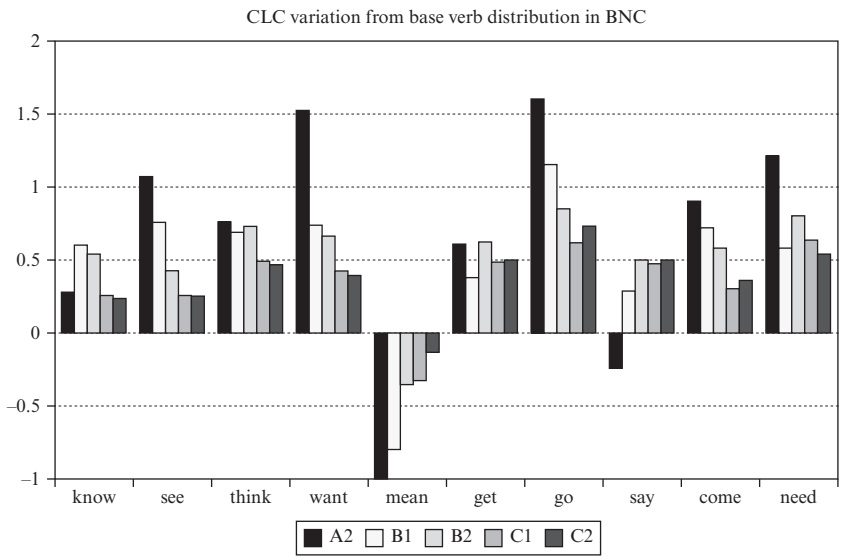
with lower frequencies, because the relevant items have been encountered more frequently (i.e. there has been greater exposure). Frequent lexical items will be overrepresented at first in L2 English, moving gradually to L1 English norms.

Ten of the most common lexical verbs in English are given in (4):

(4) know, see, think, want, mean, get, go, say, come, need

Figure 2 shows the variation (expressed as a ratio of relative frequency) in the base form of these lexical verbs in the CLC compared with the BNC, counting their present tense, uninflected forms only. Bars above the zero line indicate an over-use in the CLC compared to the BNC, while bars under the zero indicate under-use.

Figure 2 CLC variation from base verb distribution in BNC



Two findings are apparent in these data. First, these common verbs are indeed generally overrepresented in the CLC, relative to the BNC (except for the verb 'mean'), on account of the general skewing in the CLC to more frequently used items. And second, there is a general trend showing that higher levels of proficiency in the progression from A2 to C2 increasingly approximate to the BNC distribution. The bars for the higher levels are closer to the zero line.

Structural Simplicity vs. Structural Complexity

This principle is summarised in (5):

(5) Structural Simplicity (SS) vs. Structural Complexity (SC)

Structurally simple properties should be more easily acquired, in general: fewer errors, more of the relevant properties learned, etc.

Structural complexity will have the reverse effects: more errors, fewer of the relevant L2 properties learned, etc.

One consequence of (5) is that the simpler constructions of English will be acquired earlier than their more complex counterparts. For example, simpler 'subcategorisation' frames for verb co-occurrences will be acquired earlier than more complex ones.

Caroline Williams, a PhD student at RCEAL is currently analysing the 'verb co-occurrences' of English at the different levels in order to test this general prediction. In effect, she is looking at the basic construction types of English, as defined by verbs and the company they keep, in order to see whether there is a clear progression from A2 to C2 in the order in which these constructions are learned. Some simple verb co-occurrence frames of English, involving few phrases and structural relations with the verb, are shown in (6):

(6)	NP – V	*he went*
	NP – V – Part	[*the boy*] *ran away*
	NP – V – NP	*she loved* [*her husband*]

Some more complex verb co-occurrence frames are given in (7):

(7)	NP – V – Part – NP	*he looked up [the address]*
	NP – V – NP – PP	*he added [the flowers] [to the bouquet]*
	NP – V – NP – NP	*she asked him [his name]*

And some even more complex co-occurrence frames are shown in (8):

(8)	NP – V – S (Wh-move)	*he asked* [*how she did it*]
	NP – V – PP – S	*they admitted* [*to the authorities*] [*that they had entered illegally*]
	NP – V – P – Ving – NP	*they failed in attempting* [*the climb*]

Williams has found that there is a clear progression in the data from A2 to B2 in the appearance of new verb co-occurrence frames. This is shown in Tables 1, 2 and 3. Table 1 gives the verb co-occurrence frames present at A2, Table 2 gives the new verb co-occurrence frames found in B1, and Table 3 gives the new co-occurrence frames in B2.

Perhaps remarkably, Williams has found no evidence for new verb co-occurrence frames at the C levels. In other words, it appears that these basic constructions of English have been learned by B2. This confirms what others have noticed about the C levels, namely that they require a different kind, and a more subtle kind, of analysis in order to capture progress at this level. Projects are currently underway within the overall EPP that attempt to do this (www.englishprofile.org/research.html).

The progression from A2 to B2 correlates with the frequencies of these

Table 1 A2 verb co-occurrence frames

NP-V	*He went*
NP-V (reciprocal Subj)	*They met*
NP-V-PP	*They apologized* [*to him*]
NP-V-NP	*He loved her*
NP-V-Part-NP	*She looked up* [*the number*]
NP-V-NP-Part	*She looked* [*the number*] *up*
NP-V-NP-PP	*She added* [*the flowers*] [*to the bouquet*]
NP-V-NP-PP (P=*for*)	*She bought* [*a book*] [*for him*]
NP-V-V(+*ing*)	*His hair needs combing*
NP-V-VPinfinitival (Subj Control)	*I wanted to play*
NP-V-S	*They thought* [*that he was always late*]

Table 2 New B1 verb co-occurrence frames

NP-V-NP-NP	*She asked him* [*his name*]
NP-V-Part	*She gave up*
NP-V-VPinfin (Wh-move)	*He explained* [*how to do it*]
NP-V-NP-V(+*ing*) (Obj Control)	*I caught him stealing*
NP-V-NP-PP (P=*to*) (Subtype: Dative Movement)	*He gave* [*a big kiss*] [*to his mother*]
NP-V-NP-(*to be*)-NP (Subj to Obj Raising)	*I found him* [*to be*] *a good doctor*
NP-V-NP-Vpastpart (V=passive) (Obj Control)	*He wanted* [*the children*] *found*
NP-V-P-V*ing*-NP (V=+*ing*) (Subj Control)	*They failed in attempting the climb*
NP-V-Part-NP-PP	*I separated out* [*the three boys*] [*from the crowd*]
NP-V-NP-Part-PP	*I separated* [*the three boys*] *out* [*from the crowd*]
NP-V-S (Wh-move)	*He asked* [*how she did it*]
NP-V-PP-S	*They admitted* [*to the authorities*] [*that they had entered illegally*]
NP-V-S (*whether* = Wh-move)	*He asked* [*whether he should come*]
NP-V-P-S (*whether* = *Wh*-move)	*He thought about* [*whether he wanted to go*]

Table 3 New B2 verb co-occurrence frames

NP-V-NP-AdjP (Obj Control)	*He painted* [*the car*] *red*
NP-V-NP-*as*-NP (Obj Control)	*I sent him as* [*a messenger*]
NP-V-NP-S	*He told* [*the audience*] [*that he was leaving*]
NP-V-P-NP-V (+*ing*)(Obj Control)	*They worried about him drinking*
NP-V-P-VPinfin (Wh-move)(Subj Control)	*He thought about* [*what to do*]
NP-V-S (Wh-move)	*He asked* [*what he should do*]
NP-V-Part-VPinfin (Subj Control)	*He set out to win*

co-occurrence frames in the BNC, i.e. learners are first learning the more frequent frames used by English native speakers and then progressively less frequent frames, as predicted by the Frequency vs. Infrequency principle in (3). This is shown in Table 4, which gives the average token frequencies for the verb co-occurrences of Tables 1, 2 and 3, and also their average

frequency ranking, in a number of corpora including the BNC (relevant data are extracted from the VALEX lexicon, see www.cl.cam.ac.uk/~alk23/subcat/lexicon.html, which has been acquired automatically from five large corpora, both British and American, and the Web). Learners at A2 are clearly responding to the very high frequencies of the relevant co-occurrence frames in English usage that they have mastered at this point. And learners at B1 are using co-occurrence frames with higher frequencies than those at B2.

Table 4 Frequencies for verb co-occurrence frames in English corpora (including BNC)

Average Token Frequencies in the BNC etc. for the Verb Co-occurrence Frames appearing at each learner level		
A2	**B1**	**B2/C1/C2**
1,041,634	38,174	27,615
Average Frequency Ranking in the BNC etc. for the Verb Co-occurrence Frames appearing at each learner level		
A2	**B1**	**B2/C1/C2**
8.2	38.6	55.6

This progression also correlates with the increasing complexity of the structures involved. Structural complexity and frequency are generally inversely correlated in language use, i.e. the more complex a structure is, the less frequently it is used in general (see Hawkins 2004 and Wasow 2002 for relevant data, and Diessel 2004 for discussion of the relationship between frequency and complexity and their impact on language acquisition). Matching this, learners of English first learn the simpler co-occurrence frames of English before they learn more complex ones. It can be shown by a variety of complexity metrics that the new verb co-occurrence frames at each learner level are greater or equal in structural complexity compared with each lower CEFR level, and detailed research is currently in progress to document this.

Maximise Transfer

This principle, which is supported by extensive evidence for language transfer in second language acquisition (SLA) conveniently summarised in Odlin (2005), is given in (9):

(9) Maximise Transfer (MT)

> Learners prefer to maximise the transfer of L1 properties into L2. Specifically, the more similar L1 and L2 are in some grammatical/lexical domain D, the easier D is to acquire in L2, in general, since properties of L1 can be readily transferred into the corresponding properties of L2; the more different, the harder D is to acquire.

> Similar L1/L2 properties will result in fewer errors, more of the relevant L2 properties learned, and earlier acquisition.
>
> Dissimilar L1/L2 properties will result in fewer of the relevant L2 properties learned and later acquisition. With respect to errors, dissimilar L1/L2 properties will result either in more errors or in structural avoidance (and hence possibly in fewer errors, see Schachter 1974): the more obligatory or unavoidable the grammatical/lexical area in question, the more we will see errors rather than avoidance.

MT predicts that speakers of languages that have definite and indefinite articles will find it easier to acquire the article system of English, in general, than will speakers of first languages without articles. A number of second language acquisition studies have investigated this and found it to be the case: see Master (1987) and Díez-Bedmar and Papp (2008) for a comparative literature review as well as corpus data comparing Chinese learners with Spanish learners of English. In the present context notice that errors involving missing definite and indefinite articles in the L2 English of the CLC are consistently low when the L1s also have articles. This could be established by using the 'MD' or 'Missing Determiner' error code (recall the ungrammatical '*I spoke to President' versus the grammatical 'I spoke to the President' and '*I have car' versus 'I have a car' in (2) above), and by comparing first languages that have articles with those that do not.

Table 5 shows missing determiner error rates for 'the' and 'a' at proficiency levels A2–C2 for French, German and Spanish as first languages. All three languages have an article system not unlike that of English. The figures indicate the percentage of errors with respect to the total number of correct uses. For instance a percentage of 10.0 indicates that a determiner was omitted one in every 10 times that it should have appeared.

Table 5 Missing determiner error rates for L1s with articles

Missing '*the*'					
	A2	**B1**	**B2**	**C1**	**C2**
French	4.76	4.67	5.01	3.11	2.13
German	0.00	2.56	4.11	3.11	1.60
Spanish	3.37	3.62	4.76	3.22	2.21
Missing '*a*'					
	A2	**B1**	**B2**	**C1**	**C2**
French	6.60	4.79	6.56	4.76	3.41
German	0.89	2.90	3.83	3.62	2.02
Spanish	4.52	4.28	7.91	5.16	3.58

We see generally low error rates for the languages of Table 5, without significant deviation between levels.

Table 6 shows missing determiner error rates for 'the' and 'a' at all levels

for Turkish, Japanese, Korean, Russian and Chinese as first languages. These languages do not have an article system. There is a general linear improvement, i.e. a decline, in error rates across the levels with increasing proficiency (shown from left to right). Chinese shows an interesting inverted U-shaped progression, especially in the case of missing 'a', with significant improvement only at C2. These results are in accordance with MT in (9).

Table 6 Missing determiner error rates for L1s without articles

Missing '*the*'					
	A2	**B1**	**B2**	**C1**	**C2**
Turkish	22.06	20.75	21.32	14.44	7.56
Japanese	27.66	25.91	18.72	13.80	9.32
Korean	22.58	23.83	18.13	17.48	10.38
Russian	14.63	22.73	18.45	14.62	9.57
Chinese	12.41	9.15	9.62	12.91	4.78
Missing '*a*'					
	A2	**B1**	**B2**	**C1**	**C2**
Turkish	24.29	27.63	32.48	23.89	11.86
Japanese	35.09	34.80	24.26	27.41	15.56
Korean	35.29	42.33	30.65	32.56	22.23
Russian	21.71	30.17	26.37	20.82	12.69
Chinese	4.09	9.20	20.69	26.78	9.79

In the next section we illustrate in more detail how some of these data searches are actually conducted on the CLC and BNC.

How do we do these analyses?

At least half of the CLC has been manually annotated with error codes (cf. (2) above) by our Cambridge University Press colleagues prior to the initiation of the interdepartmental and interdisciplinary English Profile Programme research described here. Using a combination of these codes, together with meta-data about the candidate and simple string searches we can now collate error statistics for each exam level, language group, age, etc. (for an example recall Tables 5 and 6 on determiner errors). However, in order to collate accurate statistics over a wide range of lexical and grammatical features further annotation of the corpora is required. The RASP toolkit (www.informatics.susx.ac.uk/research/groups/nlp/rasp/) has thus been used to automatically annotate the CLC with part-of-speech tags, word lemmas, grammatical relations and complexity metrics.

Part-of-speech tags

Part-of-speech tags can be used to resolve lexical ambiguity. Consider the following sentences:

(10) They used to can fish in those towns. But now few people can fish in these areas.

A simple string search would give each of the lexical items 'can' and 'fish' a frequency count of 2 in these sentences. A study of lexical usage would need to distinguish between the noun and verb use of 'fish', however, and also between the modal and lexical infinitive usage of 'can'. By annotating the corpus with part-of-speech tags it is possible to make these distinctions, as shown in (11):

(11) They_PNP used_VVD to_TO0 can_VVI fish_NN2 in_PRP those_DT0 towns_NN2 ._PUN But_CJC now_AV0 few_DT0 people_NN2 can_VM0 fish_VVI in_PRP these_DT areas_NN2 ._PUN

Annotating a reference corpus (in our case the BNC) in a similar manner it is possible to construct accurate lexical frequency graphs (such as the one shown in Figure 2).

Lemmatisation

Lemmatisation allows for searches by citation form and by affix. Consider the sentence in (12):

(12) He was looking over at where the women stood.

By lemmatising the corpus this sentence is transformed into (13):

(13) He be+ed look+ing over to where the woman+s stand+ed.

Simple string searching makes it tedious to find all the uses of lexical items that exhibit irregular morphology (and is also error-prone). After lemmatisation, a simple search on 'be' will find all instances of the verb without having to list and search for all its inflectional derivations.

Grammatical relations

By annotating a corpus with grammatical relations the corpus linguist is able to investigate the relationships between constituents of the sentence. RASP's grammatical relations are theory-neutral, binary relations between lexical items. They may be expressed as: (|relation-type| |head| |dependant|).

For example, RASP would annotate the sentence (14) with the grammatical relations shown in (15):

(14) She was eating an apple.

(15) (|subject| |eating| |She| _)
(|auxiliary| |eating| |was|)
(|direct-object| |eating| |apple|)
(|determiner| |apple| |an|)

Once such annotation has been provided, many detailed grammatical investigations become an efficient possibility. Expanding from the example above, imagine we now wish to extract all of the direct objects of 'eating' from the corpus. Traditionally we would have to do a string search for 'eating' and then manually check all of the returned concordances for the direct objects. This is tedious at best, and at worse we might find ourselves in a situation where the direct object lies outside the concordance window (e.g. in a sentence where the direct object is displaced such as in 'The boy is eating, or perhaps he would be better described as scoffing, a juicy red apple'). By collating over grammatical relations, a frequency list can be quickly constructed, as shown for instance in (16):

(16) Grammatical relation → frequency count
(|direct-object| |eating| |apple|) → x
(|direct-object| |eating| |pizza|) → y
. . ..
(|direct-object| |eating| |words|) → z

With respect to the English Profile analysis for this specific example, we should expect to find the more frequent literal uses of 'eating' to be present at all proficiency levels whereas the abstract usage ('eating words') would occur primarily at the higher C levels.

Beyond this simple word sense demonstration, the grammatical relations can also be used to identify particular grammatical constructions (for instance the verb subcategorisation constructions identified above). To illustrate how this is done consider the following problem: *how do we find all ditransitive verbs in the corpus?* From experience we know that the ditransitive verb 'give' can occur in the structures shown in (17):

(17) Simon gives the book to her.
Toby gives the flowers to his mum.

Using these forms as a template we could search in a suitably annotated corpus for the general pattern given in (18):

(18) SUBJECT gave DIRECTOBJECT to INDIRECTOBJECT

However, this search would fail to return the sentence 'Francis gave Jodie the job' and also any sentence containing a ditransitive with a lexical item that we haven't previously identified, e.g. the one in (19):

(19) Lucy passed her sister the butter.

The solution is to search within the grammatical relations annotation for the underspecified pattern that expresses the ditransitive verb frame, as illustrated in (20).

(20) IF the grammatical relations for a sentence match:
(|subject| ?x ?a _)
(|direct-object| ?x ?b)
(|indirect-object| ?x ?c)
where ?x, ?a, ?b, ?c are variables
AND ?x is a verb
THEN the verb frame for ?x is DITRANSITIVE

Complexity measures

Finally the corpus is annotated with complexity measures that can be searched over and/or used for normalising data counts. Measures of complexity range from simple word counting of clause length to more sophisticated parsing related statistics in terms of: the total number of parses for a sentence; the average ambiguity (the number of parses divided by the number of words); average number of subanalyses; and average size of packed nodes.

In summary, by using the four automatically added annotation types described above in combination with the original error codes it is possible to construct very powerful search queries over the corpora with the result that detailed qualitative analyses can be carried out of relevance to the project's goals that test our predictions.

Towards criterial features

Our ultimate research goal is the search for criterial features and transfer effects at each of the CEFR proficiency levels, for different sets of L1s. This search is being guided by the kinds of patterns, principles and computational techniques that have been exemplified in this paper, which are in turn being informed by a wide range of computational, psycholinguistic, grammatical and cross-linguistic theories and research methods. Ultimately we will have collected numerous word frequency data (like those in Figure 2), structural properties (like those in Tables 1–3), and error counts, on the basis of which we can then enumerate the intercorrelated properties that define different proficiency levels, i.e. their 'criterial' features. These features may hold regardless of the L1 of the learner, or they may be 'L1-specific'. For example, speakers of languages without definite and indefinite articles exhibit characteristic error patterns that are different from those speaking languages with articles (see Tables 5 and 6).

This kind of study, searching for correlations among a large set of diverse lexical and grammatical properties in the progression from beginning to advanced stages of second language acquisition (SLA), has not been attempted before, and for good reason. Theories of SLA are not currently

able to make predictions for the 'horizontal' correlates for each of a set of learning stages. (For an excellent summary of the state of the art in SLA, see Doughty and Long, Eds, 2005.) A theory built on horizontal principles and predictions requires a large empirical database on the basis of which initial hypotheses can be proposed and a search capability of the type exemplified here for testing these. Such a database has not been available until recently. Acquisition theories have instead focused on the vertical or diachronic dimension, using data that have been more readily collectable hitherto, tracking just a single property (like the acquisition of the definite article) or a small cluster of properties (lexical items of a certain type or semantic class). A large empirical database can contribute to a theory with more horizontal principles and predictions, therefore, and data from a wide range of linguistic properties can now be collected, of the kind we have illustrated, on the basis of which these horizontal principles can be formulated. These in turn will enable us to define the criterial features at each level and the role of transfer from different L1 types.

Increasing proficiency levels manifest a robust and generally increasing exploitation of the properties of English, with decreasing errors by C2 and often with decreasing errors at each higher proficiency level. This can be seen clearly in Tables 1, 2, 3 and 6. More complex developmental progressions of the Chinese type in Table 6 resemble an inverted U: errors are low at first, then increase, then decline again by C2. This kind of learning curve has been found for certain types of properties and first languages in our study. Learners appear to apply a more item-based learning strategy initially, before developing a productive rule whose specific properties are mastered incompletely at first, resulting in more errors, with improved mastery later. Many other linguistic properties, by contrast, exhibit a simpler linear progression through the proficiency levels, with more consistent and steady improvements. It is an interesting and challenging research question to try and distinguish the two types of learning progressions and to predict which types of grammatical and lexical properties, and which types of L1, will result in the one or the other. A prerequisite for even formulating and addressing this question is a database of comparably tagged and parsed learner and native speaker corpora, of the type used in this study, on the basis of which the increasing exploitation of English by learners can be measured precisely and empirically. (We are grateful to John Trim for discussion of this general point after our plenary talk at the conference.)

Conclusions and practical applications

The overall patterns and principles discussed in this paper are summarised in Table 7.

Criterial features of the different proficiency levels have been exemplified for structural properties in Tables 1–3, and for certain lexical frequencies in

Figure 2, while L1-specific criterial features have been illustrated in Tables 5–6.

Table 7 Summary chart

	L2 Errors	L2 Properties Learned	Time Course of Acquisition
Frequency	fewer	more	earlier
Struct Simplicity	fewer	more	earlier
MT: similar	fewer	more	earlier
Infrequency	more	fewer	later
Struct Complexity	more	fewer	later
MT: different	more or fewer [avoidance]	fewer	later

At a technical level we can now extend the work described here to perform tasks that were thought impossible hitherto. For example, PhD student Oeistein Andersen of the Cambridge Computer Laboratory has devised an automatic system for recognising errors in texts of learner English with a very high level of accuracy indeed. Such automatic error coding can be applied to the new corpus data that we are now collecting and it will save us years of manual classification.

The collection of new data is, in fact, a high priority for the EPP at the time of writing. The CLC, despite its impressive size, is still not large enough and there are too many variables (numerous L1s, and numerous linguistic properties) with the result that some data cells are too small for significance tests. The CLC corpus is also limited in the type of data that it has, mainly answers to questions.

Let us end by outlining some practical applications of the theoretical research described here. Once we have identified and defined the criterial properties of English and transfer effects at the different proficiency levels, we can apply these findings to teaching, testing and publishing in novel ways.

With respect to teaching, we will be able to calibrate materials and syllabuses with much greater precision to the grammatical and lexical properties of English that are characteristic of each CEFR level and of the next attainable stage in learning. This research also provides the content for 'foreign market-specific' teaching materials for English targeting e.g. China and the Spanish-speaking world. Through this detailed empirical study we can legitimately claim that we have quantifiable evidence for learner errors and developmental sequences that are characteristic of these different groups of learners. Teaching materials and methods can then highlight the grammatical and lexical properties that are best presented at the different levels to Chinese versus Spanish learners, etc.

For the testing of English this research provides new content that can help to validate the scores that practitioners, i.e. examiners of English, provide. The assignment of a level and a grade to a sample of learner English relies on judgements that examiners have built up over several years of experience. Examiners 'know' what to look for on the basis of this experience, and they show high levels of agreement among themselves in the scores they assign. The empirically based work we have illustrated is beginning to describe the properties of English that examiners have come to regard as 'criterial' for each of the levels and that underlies their practical level assignments and scores.

There are also practical benefits for publishing. New publishing materials will need to be written, for Chinese and Japanese and Spanish and German markets that capitalize on this research and that present the English language in a way that is tailored to their special needs.

Finally, the research programme and the work described in this paper would not have been possible without generous funding from Cambridge Assessment and Cambridge University Press, whose support we gratefully acknowledge.

References

Briscoe, E, Carroll J and Watson R (2006) *The Second Release of the RASP System*, in Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment,* Cambridge: Cambridge University Press.

Diessel, H (2004) *The Acquisition of Complex Sentences*, Cambridge: Cambridge University Press.

Díez-Bedmar, M B and Papp, S (2008) The use of the English article system by Chinese and Spanish learners, in Gilquin, G, Papp, S and Díez-Bedmar, M B (Eds) *Linking up Contrastive and Learner Corpus Research*, Amsterdam-New York: Rodopi, 147–175.

Doughty, C J and Long, M H (Eds) (2005) *The Handbook of Second Language Acquisition*, Oxford: Blackwell Publishing.

Hawkins, J A (1994) *A Performance Theory of Order and Constituency*, Cambridge: Cambridge University Press.

Hawkins, J A (2004) *Efficiency and Complexity in Grammars*, Oxford: Oxford University Press.

Hawkins, J A (2009) An efficiency theory of complexity and related phenomena, in Gil, D, Sampson, G and Trudgill, P (Eds) *Complexity as an Evolving Variable*, Oxford: Oxford University Press.

Master, P (1987) A cross-linguistic interlanguage analysis of the acquisition of the English article system, PhD dissertation, UCLA.

Odlin, T (2005) Cross-linguistic influence, in Doughty and Long (Eds), 436–486.

Schachter, J (1974) An error in error analysis, *Language Learning* 24, 205–214.

Wasow, T (2002) *Postverbal Behavior*, Stanford, California: CSLI Publications.

10 Operationalising linguistic creativity

Wayne Rimmer
University of Reading

Abstract

At the highest levels of language proficiency, it is difficult to distinguish performance purely by lexical and morpho-syntactical units of description. Advanced language competence is partly characterised by the capability to play with language and formulate new patternings. As an example, it is shown how the non-standard lexical choice *to arouse your appetite* is motivated by a perceived gap in the collocational options available. This ability to select and manipulate a finite language system in order to maximise meaning potential is here defined as linguistic creativity. The significance of linguistic creativity to assessment is that it could offer a way of recognising and rewarding test takers operating at the highest end of the performance spectrum. Drawing on data from the Cambridge Learner Corpus (CLC) and the International Corpus of Learner English (ICLE), this paper provides examples, with commentary, of linguistic creativity in test and non-test situations. The discussion highlights the asymmetrical relationship between first and second language users and creativity: while native-speaker innovation is perceived as resourceful, learner innovation is considered deviant. In reality, the boundary between error and creativity is often blurred and it is highly context sensitive. It is suggested that the changing demography of English is both licensing and encouraging creativity in a second language context. The conclusion outlines the practical challenges of operationalising linguistic creativity in an environment which favours maintaining the status quo in systems of assessment.

Background: advancing language description

An adequate description of the language variation elicited during assessment is critical. Language performance varies on numerous dimensions so assessment must isolate and describe the features that are relevant to the underlying construct. This process becomes progressively more challenging as learners develop their linguistic competence, because of the increasing extent and complexity of the language sample which is produced. This paper argues

that language variation should be understood differently as learners progress towards proficiency. Drawing on corpus data from an empirical study, it makes a case for linguistic creativity as a factor which allows meaningful differentiation between test takers at the higher end of the performance spectrum.

The problem of accounting for language variation is particularly acute with advanced learners, advanced understood loosely as covering the C levels of the CEFR. There has been a general lack of provision for the needs of high-level learners and paucity of research in this area (c.f. Leaver and Shekhtman 2002). In particular, there is much uncertainty about the grammar syllabus for advanced learners. A common position is that the main areas of grammar have been covered before the advanced level: Shaw and Weir (2007) in explaining how grammar is treated in the Cambridge ESOL exams state that, 'Above FCE [B2] level there are relatively few new aspects of grammar to introduce' (2007:111). Their comparison (2007:114) of the structural resources available to test takers at B2, C1 and C2 levels is illuminating.

> FCE: Learners at this level are able to use all the main tense forms and structural resources of English.
> CAE: The writer is able to adjust his or her writing to suit the context and target reader adopting a style that will convey the message in an appropriate way.
> CPE: The writer is able to use grammar to organise writing effectively and to express subtle differences of meaning and attitude.

It would appear that there is a transition at the C levels from learning new grammar to deploying familiar structures more skilfully. Potentially, this is a very rich distinction but it needs corroboration and exemplification with language data. Research findings from the English Profile Programme, an attempt to bolster CEFR with empirically derived descriptors for the linguistic components of the six levels (see Taylor and Barker 2008:250), supports the notion of a cut-off point in the acquisition of new grammar structures. English Profile is still at a nascent stage but work carried out to date (presented in Barker, Kurtes and Sylvester 2008) suggests that for the first time it may be possible to substantiate degrees of attainment with detail of the grammatical forms and structures integral to performance at that level. This is very encouraging but the contribution of English Profile to CEFR may be less directly applicable to the higher levels where acquisition has reached a ceiling effect. For example, the work of Hawkins and Buttery (this volume) empirically confirms that advanced learner language is not characterised by a larger repertoire of grammatical items. At C levels, learners have established a full range of grammar, therefore, as Hawkins and Buttery note, the descriptors '. . . require a different kind, and a more subtle kind, of analysis in order to capture progress at this level'.

Clearly, the acquisition of new structures is not linear and it peters out by the C levels, hence the genuine quandary of what to teach advanced learners (Hall and Foley 2004). Grammar is not everything, of course, but it is especially significant in that it represents the main organising principle in the curriculum (Nunan 1988), partly due to tradition (Brennan 2007), partly due to learner expectation (Jianbo and Greenall 2008). When grammar is missing from language description, or under-specified, there is a very big hole to fill. Advanced language competence consequently remains seriously underspecified. However, the fact that the construct is difficult to describe does not mean that it is unattainable (Black 2007). What is clear, as Hawkins and Buttery conclude above, is that at higher levels we need sharper tools of description. Defining increasing competence in terms of access to newer language simply falls short.

It is unlikely that the solution lies in better descriptors and rating instruments. Even if reliable scales were available, there is so much inherent variation in development and performance at C levels that they would understate the individualised process of language acquisition and use. It is a truism that language is context dependent (e.g. Berman and Nir-Sagiv 2004, Bright 2005, Chapelle and Douglas 2006, Purpura 2004) so generalisations about hypothetical performance in hypothetical situations will not correspond to specific usage. The factors which distinguish advanced language are too subtle to be translated into discrete criteria. Consider this opening of an essay from the International Corpus of Learner English (ICLE) (Granger, Dagneaux and Meunier 2002), a collection of advanced writing.

> My opinion is light. It is like a balloon. Once it is inflated by a refreshing breath of inspiration, it goes up in the air as high as possible.

The first two sentences mirror each other in both having a subject – copula – complement structure. The simple understated opening of the essay, underscored by the playful alliteration and assonance of *light*/*like*, creates an appropriate feeling of levity. In contrast, the finite adverbial clause in the third sentence is substantial: the verb is passive and the noun phrase is both pre- and post-modified. There is an air of expectation as we wait for the main clause and the sudden transition to more involved syntax marks the unfolding of the image. The balloon simile is hardly original but it has a deceptive resonance. The flight of the balloon denotes freedom of expression and exhilaration, but a darker side to the image is the potential danger of uncontrolled movement, for *as high as possible* can only lead to disaster. Note too the iconicity as the three adverbials in the main clause *up in the air as high as possible* increase in word length, one – three – four words, as the balloon increases its altitude. Freedom is a risky thing. The ambivalence of the simile is artfully reinforced by the simple/complex juxtaposition of sentences 1 and

2 with sentence 3. The unwary reader is lured by the first two sentences into accepting the image at face value while the artfulness of the third sentence suggests a deeper meaning.

This is a very potent and arresting beginning to an essay. The effect does not lie so much in individual linguistic features, e.g. the adverbial clause, as in the way that they interact and develop the topic. The writer selects and exploits a range of language features to create a text which transcends their individual value. No language is a perfect system so there is often a tension between what we want to express and what we can express. Infinite expression is somehow possible through finite means. This is particularly pertinent to learners as, by definition, their language resource is incomplete. The facility to manage and manipulate language for fuller and deeper self-expression is the hallmark of skilled performance. Creativity is this force that allows the user to maximise linguistic competence and construct discourse which is purposeful and semantically resonant in a concrete environment. It demands a thorough knowledge of the language system so that fresh meanings can be made and combined without violating the norms of standard usage. As such, creativity is particularly relevant to advanced learners as they are competent and experienced enough to recognise and exploit the potential and limitations of their language resource.

Creativity resists neat definition since it is essentially a holistic concept. Maybin and Swann (2007) analyse creativity along three dimensions: textual, contextualised and critical.

Table 1 Dimensions of creativity (adapted from Maybin & Swann 2007:513)

Dimension	Focus	Examples
Textual	Vocabulary, grammar, phonology	Word play, intonation, emoticons in emails
Contextualized	Language as used by participants to respond to particular sociocultural and sociohistoric contexts	Joint construction of a narrative, cultural assumptions behind jokes
Critical	Evaluation and critique of status of language and participants	Moral stance, joking that is socially subversive

Textual creativity is likely to be of most interest to language testing because the language component is most transparent. This paper accordingly focuses on creativity in a fairly traditional sense of grammar and vocabulary. Creativity is easier exemplified than defined so reference will be made to the CLC. As this is a written corpus, the methodology confines the study to creativity as a written phenomenon. This is certainly not to deny or denigrate the role of creativity in speaking, it is simply a restriction of scope imposed by the data set available.

Creativity in learner corpora

The literature on creativity very often refers to its ludic nature. Cook (2000:128) argues that language is a game we all play. Learners join in too as in this humorous excerpt from a Proficiency script in the CLC. (All corpus citations are verbatim with errors uncorrected).

> As ever, the prespective of gushing forth the convetional – "Hello Mark, it has been a long time, hey" – didn't sound very natural. He was late, once more.

The candidate describes meeting an old flame after a long separation. The essay plays between the pain and pleasure of such an encounter, a conflict of emotions enforced by the language use. The head of the noun phrase *prespective* (sic) is post-modified by the vivid *gushing forth*. There are three instances of *gush forth* in the BNC.

> . . . the water gushed forth in a vast fountain.
> . . . while the legal fees gushed forth like blood from the wounds.
> . . . a torrent of words gushed forth.

Syntactically, the BNC citations (all from fiction books) differ from the Proficiency script in that they are finite and intransitive. The learner language is very striking in taking direct speech as an object. Semantically, the BNC concordance lines are similar in that the outburst is damaging. In the first instance, there is a literal flood; in the second, a bitter and expensive legal wrangle; in the third, a stressful verbal confrontation. In the light of the, admittedly meagre, corpus evidence, *gushing forth* suggests a negative response to the reunion. However, the emotional charge is belied by the laconic bluntness of *He was late, once more*. The frame of time adverbials around the section, *As ever . . . once more*, also imposes some banality on the scene. Our mixed reaction to the text mirrors the writer's ambivalence. On the one hand, the experience is poignant and significant. On the other hand, it is sadly anti-climatic. The writer engineers the two conflicting interpretations through nominalisation, the heavy and graphic noun phrase which dominates the text. There is some irony in her skill in manipulation. She can influence the unknown reader but not the once dear Mark. Perhaps mastery in language compensates for failure to control real life events. The humour is bittersweet.

This excerpt was highly original in its structure and content. Creativity, however, is not exclusively about novelty of expression and meaning. Carter (2007) comments that creativity operates in two competing directions. There is a centripetal force for conformity and similarity; there is a centrifugal force for dissidence and rebellion. Only the latter is associated with innovation. Familiar language in a new context may be equally evocative. Below is the beginning of another Proficiency script.

> When I was boy, I used to spend my holidays in my grandmother's house. It's an old farm, standing in the middle of the fields. Not far from it lies an old forest. It's the kind of forest you can only think of as old. And in that forest lies a stone, a huge stone called "the Goblin Stone".

The opening adverbial of time and the semi-modal *used to* plant the narrative firmly in the past. It is somewhat disconcerting then when the essay switches into the present tense. Is this story about the boy or about the man? It is as if the writer has become so engrossed in the recollection that the past becomes immediate to him. The three-fold repetition of *old* reminds us that although the story is told in the present it concerns the past. There is also repetition of *lies*. The repetition is enforced by the syntactic parallelism of the two occurrences: prepositional phrase + *lies* + post-verbal subject. *Lies* is of course ambiguous, as in the Larkin poem *Lying in bed* (Larkin 1986). Forests and stones cannot literally lie but they can deceive observers by not being what they seem. The stone which sealed the tomb of Jesus was not the unsurpassable obstacle that the Jews intended (Matthew 28:66). The candidate tells the story so simply and directly that we accept its truth value. *Lies* is an uncomfortable indication of the possibility of misinterpretation and the fallibility of memory. The ambiguity is perhaps a warning for the tone indeed changes in the last sentence.

> And in that forest lies a stone, a huge stone called "the Goblin Stone".

The grandmother suggested love and comfort. The stone introduces a sinister element. We also move from the real to the fantastic, although *lies* questions the validity of this transition. The opening *And in that forest* is strangely arresting. The initial *And* is certainly not deviant. Burchfield (1998:52) notes that there is a long history of initial *And* from prestigious writers including Shakespeare. A search for *and in that* (+ noun) in the BNC (British National Corpus) is illuminating.

> I dreamed – and in that dream Frankenstein was born.
> And in that twilight world between waking and sleeping she saw that it was a face she knew, her own.
> And in that flare and roar of lightning, Cardiff saw a hideous face looking at him.

All three citations depict highly charged and traumatic events. *And in that . . .* is often associated with genres of fantasy, horror and science-fiction. The candidate uses the phrase to change the direction of the narrative. The syntax is then rearranged to give *stone* significance and awe.

> . . . a stone, a huge stone called "the Goblin Stone".

The first mention of *stone* is as an unmodified noun. The critical information about its size and name is marked off with a comma and separated in an appositional supplement. This noun phrase is marked as the focus of the essay. *Huge stone* is post-modified by an *-ed* participle clause, . . . *called "the Goblin Stone"*, rather than a finite relative clause, . . . *which is/was called "the Goblin Stone"*. The non-finite clause allows the writer to obscure the time reference, *is* or *was called*, so the status of the temporal location of this narrative is obviated. The simplicity of this text is beguiling. Grammar, vocabulary, even punctuation, all combine to give the language a disturbing undercurrent of suspense and tension. The creativity is in weaving these different elements together and unleashing them on the reader to maximum effect.

Discussion

Creativity challenges test takers to take language to its limits, a zone where error and innovation become difficult to distinguish. Consider this culinary example from the CLC.

> . . . it won't, I dare hope, fail to arouse your appetite and curiosity in Russian food.

Arouse takes a co-ordinated noun phrase but its collocational relationship with each noun is unequal. Two corpus-based collocation dictionaries (Hill and Lewis 1997, Runcie 2002) allow *arouse curiosity* but not *arouse appetite*. The latter is attested twice in the BNC but not often enough to be registered there as a collocation so it is certainly an unusual phrasing, which if not inappropriate is at least highly marked. If we accept the weight of corpus evidence in regarding *arouse appetite* as non-standard, following the tradition whereby error is defined by corrigibility (Corder 1984:163), we should consider a more viable formulation. The following verbs are identified by the BNC as collocates of *appetite* in the sense of 'make more':

Table 2 Collocates of appetite

increase sharpen stimulate whet work up	appetite

A test of their aptness is to substitute them in the context. However, when we do this they all seem inadequate. *Increase* is somewhat bland; *sharpen* introduces an incongruent military metaphor; *stimulate* is formal; *whet* is too weak; *work up* is associated with vigorous exercise. None of them seduce

the epicurean side of the reader. At this point, the resourceful test taker may search the lexicon for viable alternatives. *Arouse* collocates with *curiosity* and it also means increase.

Table 3 Collocates of *arouse*

arouse	suspicion passion hostility controversy interest

Arouse is a loaded word, it seldom occurs in a neutral context. *Passion* is revealing for *arouse* often carries sexual connotations: *man* is also a collocate. The very phrase *sexual appetite* attests to the tight connection between food and sex. This mixture of strength and sexual suggestiveness makes *arouse* very attractive for the food text. Indeed, going back to the BNC, one of the only two citations for *arouse appetite* occurs in a sexually suggestive context.

> The spectacle of a beautiful boy in a dress and dreadlocks confiding to the nation's media that he preferred "a cup of tea" to sex appeared both to arouse the public appetite for sexual frisson and deflate it with humour, honesty and a curious sort of innocence.

The phrase also stands out here for it is highly stylised. The language is meant to be as striking as the personality (the erstwhile pop star, Boy George) described. Not only is the collocation colourful and amusing, it is iconic, for it mirrors the pop star's alluring mixture of the controversial and conventional by being on the edge of linguistic acceptability itself. If we allow this usage from the BNC as skilled writing, we should also recognise the achievement of its counterpart in the CLC. *Arouse appetite* is not a standard collocation but its effectiveness for the candidate in this particular context is maximal. The process of searching, rejecting and substituting collocates is a creative process for which the candidate should surely be credited rather than penalised.

Despite this, genuine error through lack of competence is a reality of even advanced language use. Not every deviation can be counted as creativity. Cross and Papp (2008) in their contrastive analysis of three learner corpora consider innovation to be legitimate only if there is evidence of a stable language competence spawning that specific usage. Cross and Papp examined verb + noun combinations in Greek, German and Chinese learners in the ICLE. They found not only that Chinese writers produced many more errors but that they were less likely to experiment with alternative verb or noun collocates, e.g. **carry out an exam* (my example), suggesting to the authors that Chinese learners, for a variety of linguistic and cultural reasons,

are not as creative as their European counterparts. This led Cross and Papp to consider the conditions for innovation: what determines whether *carry out an exam* is error or innovation? They distinguish between convergent, divergent and incomplete knowledge based on native speaker norms. Convergent usage is well attested in native-speaker corpora so it is impeccable. Divergent usage is idiosyncratic but motivated by a sound knowledge of the target language. For example, the phrasal verb *carry out* is semantically associated with unpleasant and arduous tasks. While there is no instance of *carry out an exam* in the BNC, there is *carry out ~ assessment, audit, evaluation* and *test.* Exams certainly fit into this category so the sequence *carry out an exam* is more acceptable than, say, **carry out a picnic*. The latter would be an example of incomplete knowledge, basically arbitrary language use not founded on a deep understanding of the lexicon and grammar. Cross and Papp argue that the case for innovation can be built up by reference to the fullness of the learner's competence.

> We need to be able to measure creativity and define its limits. As long as new innovative collocations are structurally, semantically and pragmatically motivated and justified, that is they are entrenched in old conventional patterns and are recognized as creative uses of the language that are produced for some special purpose (for instance for purposes of informality, irony, humour. . .), they can be accepted as legitimate combinations. However, when there is no such feeling that the non-conventional usage was based on conventional patterns and produced for a particular purpose, it will always be conceived as a learner error and will serve as a marker of non-nativeness and taken as evidence of lack of control that the learner has over the L2 (2008: 76).

Here the demarcation zones between creativity and error are a little too firmly drawn. In practice, much usage will occupy an uncomfortable point on a cline between divergent and incomplete knowledge. For example, from the quotation above, the metaphorical use of *entrenched* is intriguing.

> . . . that is they are entrenched in old conventional patterns . . .

For me, *entrenched* has negative connotations, as the etymology from the root *trench* suggests, while the meaning of the *that*-supplement clause above is unequivocally positive for it defends creativity in given contexts. The BNC confirms my intuition for most of the citations of *entrenched* as an adjective and past participle are negatively marked.

> . . . Johnny's double standards, and his entrenched belief in the superiority of the male.
> . . . the Blues once again find themselves entrenched in a fight for survival.

> On the one hand they resented the entrenched power of the landed aristocracy.

Entrenched often suggests stubbornness and an unwillingness to yield which is surely very far from the authors' concept of creativity: innovation is exactly the opposite force, one dependent on flexibility and shifting values and boundaries. Their use of *entrenched* is not natural for me nor is it supported by appeal to expert-user corpora. Of course, this is a published academic book about linguistics so, as with poetry, it is somehow beyond the rules of this particular discourse to challenge the language. The point of the *entrenched* exercise is to show that even paradigms of language competence, linguists, yield language samples which are not impervious to criticism. Because of this, the native-speaker corpus should not be treated as the ultimate arbitrator between error and innovation.

For too long, creativity has been regarded as the province of the native-speaker. Unconventional learner language tends to be written off as error (Prodromou 2007). As Cross and Papp point out, this ignores the crucial dimension of context of use. Furthermore, it does not take account of the pressures mounting against a monolithic model of English. Graddol (2006:101) predicts that in 2010 two billion people will be learning English. The ownership of English can no longer be restricted to Anglophone countries and second language learners will have an increasing influence on the spread and development of English. Jenkins (2006) argues that examination boards must accommodate for the consequent diversity of standards. The impact of the globalisation, some would argue fragmentation, of English is too complex to rationalise but it will surely increase the arsenal of English and provide users with more language options to choose from. Creativity can only thrive in this new environment of diversity, linguistic richness and tolerance.

Conclusions: language form is only part of forming language

This paper has presented evidence of second language creativity by test candidates. It has also argued that creativity makes a valuable contribution to construct validity at advanced levels. However, it has not demonstrated that creativity is measurable. These final words show the challenge of operationalising the construct.

More open and affective tasks

To exercise creativity, test takers need an environment which encourages extended language use and a degree of freedom. Discrete test items like

multiple-choice clearly do not do this. Complexity is most abundant when there is evidence of a high degree of engagement with the task. Mechanical, routine tasks will generate responses of the same ilk. Careful task design is therefore crucial. Because affect is by definition very individual, it is difficult for large exam boards to create writing opportunities which appeal to a wide candidature. One alternative is to tailor exams to smaller groups based on nationality, age-group, interest, or even a combination of these. The other is to give test takers a choice of task. The latter is more practical but it will affect comparability of responses and thus depress reliability (Hughes 1989:38).

Generally, there does need to be some reevaluation of reliability in measuring creativity for it involves a highly personalised response to the text. Inevitably, a plurality of interpretations poses a reliability problem in assessment. Yu (2007) comments that the postmodern insistence on the legitimacy of any textual interpretation presents testers with a dilemma akin to the classic reliability vs validity trade-off. The construct ultimately resides in the interaction between the text and the rater, a process which cannot be pre-specified or perhaps even verbalised. As such, the emphasis should be on training raters to recognise complexity, not on scale validation. Raters need thorough exposure to scripts displaying a range of linguistic features that contribute to creativity. Rating instruments will not be rendered redundant but they will become reference tools and lose their sole authoritativeness. This does put more responsibility on the individual rater, and humans are fallible, again a reliability issue. Still, it is unlikely that creativity will be the only construct of interest in a test. Other more objective measures can balance out the reliability of the test as a whole. Rater training and descriptive statistics can also minimise the threat to reliability.

An additional challenge in high-stakes tests is the pressure for test taker conformity and conservativeness. If a major lifestyle decision rides on a test score, candidates are likely to play safe with language and avoid creativity if it involves risk-taking. This is a validity problem because a false picture of a test taker's ability will be projected. Arguably, it is also an ethical concern as the test is discouraging candidates from natural expression of their thoughts and beliefs. At lower levels, where the language demands and expectations of performance decrease, the play-safe factor, so to speak, is neutralised. Ideally at C levels, exam boards need to gather a wider sample of language and turn to other assessment methods such as portfolios. Unfortunately, this adds a longitudinal dimension to the process which is probably at odds with stakeholders' expectations of swift score reporting.

Human rating essential at higher levels of performance

Educational Testing Service and, to a lesser extent, Cambridge ESOL are researching tools which give an automatic evaluation of a writing script.

(Machine reading of discrete item tests is commonplace, of course.) The mechanism is based on counting units and comparing the result to a norm. Thus, mean sentence length could be indicative of performance. The advantages of machine rating are logistical, savings of time and money, resulting in cheaper tests with faster score-reporting. However, present technology is a long way from developing instruments which are sensitive enough to detect and evaluate linguistic creativity. The human element in the process can probably never be eliminated. Sentences cannot be parsed for style, originality, response to context, or any of the factors which combine to produce creativity. That is not to say that automatic marking has no applications at C levels. Scripts could be machine parsed as a preliminary for human rating. This would save the rater time and generate comparative statistics across a body of candidates. There is no reason why machines could not assign a score for areas such as spelling, punctuation and perhaps grammatical accuracy. Indeed, Cambridge ESOL is considering a combination of machine and human rating at higher levels (Papp, personal communication). However, the assessment of creativity demands a meta-cognition that only humans can provide.

Input from learner corpora

Second language testing should be informed by authentic learner language. Barker (2008) is correct that corpora now play a larger role in assessment, an example being the English Profile Programme, but the paucity of learner corpora is an obstacle to progress. The ICLE is well-established, the second edition was released in 2008, but at 2.5 million words of written academic English it is incomparable in size and representativeness to native-speaker corpora such as the BNC. The CLC is more balanced in level and authorship and it is tagged but, largely for commercial and copyright reasons (Taylor, personal communication), it is not accessible to the general public. Creativity is a complex phenomenon which will be better understood and appreciated when there is more learner data available for analysis.

Just as there is more to assessment than counting and comparing, there is more to language competence than grammar and vocabulary. Creativity is dependent on a highly developed language resource but it is not simply the product of specific language forms, even where they are marked by low-frequency, or morpho-syntactic complexity. There is no direct link between syllabus items and creativity. Test specifications cannot reduce the process to a checklist of desirable linguistic features. First and foremost, creativity celebrates language choice and freedom. These are not traits which can be determined *a priori*. Thus, we are also faced with a choice. As discussed, we can dismiss creativity as a threat to reliability and the comfortable status quo. Alternatively, we can interpret creativity as a liberating and essentially

human characteristic which can transform the assessment experience. Learner corpora show that creativity is a real phenomenon in a second language. The direction assessment chooses will depend much on how seriously it takes its commitment to corpus-based language description.

References

Barker, F (2008) *Using corpora for language assessment: trends and prospects*, paper presented at the ALTE conference, Cambridge, UK.

Barker, F, Kurtes, S and Sylvester, K (2008) *Research Notes* 33, Cambridge: Cambridge ESOL.

Berman, R and Nir-Sagiv B (2004) Linguistic indicators of inter-genre differentiation in later language development, *Journal of Child Language* 31, 339–380.

Black, B (2007) Critical thinking – a tangible construct? *Research Matters* 3, 2–4.

Brennan, B (2007) Review of Essential Business Grammar and Practice – Elementary to Pre-intermediate, *Modern English Teacher* 16 (4), 72–73.

Bright, W (2005) Contextualizing a grammar, *Studies in Language* 30 (2), 245–252.

Burchfield, R (1998) *Fowler's modern English usage,* Oxford: Oxford University Press.

Carter, R (2007) Response to special issue of applied linguistics devoted to language creativity in everyday contexts, *Applied Linguistics* 28 (4), 597– 608.

Chapelle, C and Douglas, D (2006) *Assessing language through computer technology*, Cambridge: Cambridge University Press.

Cook, G (2000) *Language play, language learning,* Oxford: Oxford University Press.

Corder, S (1984) Idiosyncratic dialects and error analysis, in Richards, J (Ed.) *Error analysis*, Essex: Longman, 158–171.

Cross, J and Papp, S (2008) Creativity in the use of verb + noun combinations by Chinese learners of English, in Gilquin, G and Papp, S (Eds) *Linking up contrastive and learner corpus research*, "Language and Computers – Studies in Practical Linguistics" Series, Volume 66. Amsterdam, New York: Rodopi, 57–81.

Graddol, D (2006) *English next,* Manchester: British Council.

Granger, S, Dagneaux, E and Meunier, F (2002) *International corpus of learner English,* Louvain, Belgium: UCL Presses Universitaires de Louvain.

Hall, D and Foley, M (2004) Advancing the advanced, in Pulverness, A (Ed.) *IATEFL 2003 conference selections*, Kent: IATEFL, 44–46.

Hill, J and Lewis, M (1997) *LTP dictionary of selected collocations,* Hove: Language Teaching Publications.

Hughes, A (1989) *Testing for language teachers*, Cambridge: Cambridge University Press.

Jenkins, J (2006) The spread of EIL: a testing time for testers, *ELT Journal* 60 (1), 42–50.

Jianbo, W and Greenall, S (2008) Lessons to learn from ELT in China, in Beaven, B (Ed.) *IATEFL 2007 conference selections*, Kent: IATEFL, 76–86.

Larkin, P (1986) *The Whitsun Weddings,* London: Faber & Faber.

Leaver, B and Shekhtman, B (2002) *Developing professional-level language proficiency*, Cambridge: Cambridge University Press.

Leech, G, Rayson, P and Wilson, A (2001) *Word frequencies in written and spoken English*, Harlow: Pearson Education Limited.
Maybin, J and Swann, J (2007) Everyday creativity in language: textuality, contextuality, and critique, *Applied Linguistics* 28 (4), 497–517.
Nunan, D (1988) *Syllabus design,* Oxford: Oxford University Press.
Prodromou, L (2007) Bumping into creative idiomaticity, *English Today* 89 (23), 14–19.
Purpura, J (2004) *Assessing grammar,* Cambridge: Cambridge University Press.
Runcie, M (2002) (Ed.) *Oxford collocations dictionary for students of English,* Oxford: Oxford University Press.
Shaw, S and Weir, C (2007) *Examining writing: research and practice in assessing second language writing*, Cambridge: UCLES/Cambridge University Press.
Taylor, L and Barker, F (2008) Using corpora for language assessment, in Shohamy, E and Hornberger, N (Eds) *Encyclopaedia of Language and Education* (*2nd ed)*, Springer Science + Business Media LLC, 241–254.
Yu, G (2007) Students' voices in the evaluation of their written summaries: empowerment and democracy for test takers? *Language Testing* 24 (4), 539–572.

11 The consequences of examining through an unfamiliar language of instruction and its impact for school-age learners in Sub-Saharan African school systems

Pauline Rea-Dickins, Guoxing Yu and Oksana Afitska
University of Bristol, UK

Abstract

A significant number of children in Sub-Saharan Africa demonstrate their learning in formal school examinations through the medium of English and the function of English as a mediating tool for subject learning has become increasingly controversial in recent years. One important facet in this controversy has to do with the impact that a language, that is not the first language of the majority of students, has in determining their progression within an education system as well as their educational outcomes. The context for this review relates to the role of language as a critical factor for effective learning in the African context and it examines issues of test fairness and social consequences through the lens of the individual, analysing the evidence available on: (i) the necessity and the complexity of providing test accommodations for additional language learners and how linguistic accommodations may affect learner performance in formal assessment settings; and, given the importance of effective classroom teaching and learning to prepare students for examinations, (ii) the effects of classroom language(s) use on learner engagement in subject classrooms and achievement in formal examinations.

Introduction

A significant number of children in Sub-Saharan Africa (SSA) demonstrate their learning in formal school examinations through the medium of English and the function of English as a mediating tool for subject learning has become increasingly controversial in recent years. One important facet in this controversy has to do with the impact that a language, that is not the first language of the majority of students, has in determining their progression within an education system as well as their educational outcomes. The context for this review relates to the role of language as a critical factor for effective learning, with specific reference to SSA and arises from an investigation into the impact of the language of examinations and media of instruction (Kiswahili and English) in secondary schools in Zanzibar on the examination performance of learners acquiring subject knowledge and understanding through a language that is not their first (ESRC/DfID Grant RES-167-25-0263). Empirical evidence suggests that language constitutes a determining factor in the demonstration of achievement in the formal examination of school subject knowledge and that, as examples, the use of an unfamiliar language as the medium of instruction is a factor in underachievement, in school effectiveness, that it contributes to drop out rates and grade repeating and that girls may particularly be disadvantaged. However, very little is known about why and *how* pupils may be disadvantaged in demonstrating subject learning through a second or additional language (EAL)[1], in particular, in SSA.

The quest for implementing and understanding quality education processes and provision is a universally shared goal but the policy contexts within which these aspirations reside may be quite different in so far as issues of language are concerned. The more familiar and recent charters include the *No Child Left Behind Act of 2001* (Public Law 107-110, USA) and *Every Child Matters* green paper (2003 UK). Less prominent is *World Declaration on Education for All* and *Framework for Action* with their goals for SSA aimed at addressing poverty reduction and providing quality *Education for All* (World Conference on Education for All, 1990) to meet the challenges of globalisation. Amongst some policy makers in SSA, there is a view that 'education equals English', a perception found to be pervasive amongst parents as a significant stakeholder group (Rea-Dickins, Rubagumya and Clegg 2005). English as the school medium still remains largely unquestioned in SSA (but see below) and, where this is the case, there is a similarity with the mainstream contexts in the USA and England. There is also evidence of a growing number of English 'only' schools in SSA reflected in the burgeoning private sector (e.g. Rugemalira 2005). This, however, is where similarities end. For additional language learners, English is the dominant language in the UK and USA, i.e. the majority of the population speaks English, and their

teachers are expected to be competent users of English. For students in SSA, English is an exoglossic language (i.e. not the first language of the majority of the population) and their teachers are learners of English too and are not necessarily competent users of English. It has been noted, for example:

> In Zanzibar, Kiswahili is the medium of instruction in primary schools and English at secondary and higher levels. However, there is perceptible weakness in language proficiency of teachers and students in English (Ministry of Education and Vocational Training, Zanzibar, *Education Policy*, 2006:35).
>
> Graduates of basic education are weak in both languages, the vernacular and English. They lack effective communication skills . . . (2006:2).

It is also somewhat alarming that as far back as 1953, UNESCO (1953:6) declared: 'We take it as axiomatic . . . that the best medium for teaching is the mother tongue of the pupil'. More recently, Williams (2006:187) also highlights:

> . . . that poor countries often operate expensive and often complex language policies, whereas rich countries usually operate simple and relatively cheap language policies. Thus the policy in Malawi and Zambia involves home-school language switching, with teaching in at least 2 languages, while countries such as England and France operate what is overwhelmingly a monolingual policy, in a language that is the first language of most learners and teachers.

The complex and wide ranging issues involved can be analysed from a range of perspectives, at political, social and individual levels. The focus that this review takes is an analysis of test fairness and social consequences through the lens of the individual, analysing the evidence available on: (i) the necessity and the complexity of providing test accommodations for EAL learners and how linguistic accommodations may affect learner performance in formal assessment settings; and, given the importance of effective classroom teaching and learning to prepare students for examinations; (ii) the effects of classroom language(s) use on learner engagement in subject classrooms and achievement in formal examinations.

Measuring scholastic achievement

Impact factors: necessity, complexity and discrimination

The importance of using educational assessment for monitoring the academic progress of pupils, school effectiveness and overall educational quality of school systems is both widely acknowledged and used in both developing

and developed countries to meet the challenges of globalisation and poverty reduction. However, the opportunities to use examinations as a lever for change (Kellaghan and Greaney 1992, 2004) in monitoring and improving education quality in SSA are often missed, misused or even abused, leading to a 'serious waste of scarce educational resources' (Kellaghan and Greaney 2004:13). Rather seriously, this raises issues of social and individual inequality with discrimination 'against minorities, rural populations, girls, and students whose first language differs from that of the examination' (Kellaghan and Greaney 2004:7), normally a language of Europe from colonial times based on certain functional assumptions of the nature of society (Clayton 1998), for example, to maintain the *status quo* because of political priorities of unification and modernisation, or parental pressure (Williams and Cooke 2002).

In terms of data on the actual performance of EAL learners in examinations, there is mounting evidence of considerable disadvantage. Hazel, Logan and Gallagher (1997), for example, reported that question type and the context within which examination questions were set were likely to favour male students, with boys outperforming girls on traditional forms of assessment especially multiple-choice questions (MCQ). With reference to test results in Niger, it has been demonstrated that learners who started in their mother tongue (L1) could read and write better even in the second language (Hovens 2002). Findings from Mwinsheikhe's research (1991–1995, 2001), cited in Brock-Utne (2002:15), showed that girls in particular did exceptionally badly in the sciences, with 95% failing in biology, and over 85% failing in chemistry and physics examinations. Probyn (2006), too, identified how learners' ability in the L2 may limit their capacity to show learning in L2-medium assessment formats: they may fail to understand examination questions and be unable to answer them. This appears to be especially the case with open-ended formats that require learners to produce a sentence or more (Mwinsheikhe 2002, 2003, Probyn 2006).[2]

The underachievement of students in SSA countries, as demonstrated in both national and international educational assessments, and its associated problems such as gender disparity and high drop out and grade repetition rates (Yu and Thomas 2008) raise at least three questions. The first has to do with whether problems of underachievement are related to a language of instruction that is different from the students' home language as well as their teachers (Brock-Utne and Holmarsdottir 2004). Secondly, there is the question as to whose language(s) should be used as medium of instruction (Brock-Utne 2001) to meet the targets of *Education for All* (World Conference on Education for All, 1990) and, thirdly, there is the question as to which language(s) should be used for formal and high-stakes examinations. In developed countries, such as the UK and USA, increasing numbers of students classified as EAL learners who are assessed through the medium of English at school, are also widely identified as low achievers, compared

to those whose first language is the language of assessment (e.g. Abedi and Gándara 2006 in the USA, Hargreaves 1997 in the UK). We know, however, that the language of examinations may affect the psychometric features such as the reliability and validity of the examinations (Abedi 2002, Shorrocks-Taylor and Hargreaves 1999, 2000), as stated in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education 1999:91):

> For all test takers, any test that employs language is, in part, a measure of their language skills. This is of particular concern for test takers whose first language is not the language of the test. Test use with individuals who have not sufficiently acquired the language of the test may introduce construct-irrelevant components to the test process. In such instances, test results may not reflect accurately the qualities and competencies intended to be measured.

The language of examinations, therefore, not only has implications in relation to the ethics and fairness of the examinations towards the individual students (Bailey and Butler 2004) but also for the social and political arenas of an educational system and its planning and implementation (see Greaney and Kellaghan 1995).

Language for instruction and assessment has been a politically sensitive issue in many developing countries for decades (e.g. Mlama and Matteru 1978), and highlighted more recently by Kelleghan and Greaney (2004). Given that formal large-scale educational assessments in SSA have considerably higher stakes than those in the USA and UK – for example, (i) the participation level in lower secondary education in Sub-Saharan Africa is very low: 38% (UNESCO 2007:58), with a considerable number of students leaving school for good, based on their performance in examinations delivered in English or French at the end of primary education; (ii) only 27% of pupils who begin school in South Africa exit with a school-leaving certificate after the 12th grade (Heugh 2000:30) – it is of vital importance to understand what is the impact in relation to language on examination performance. For example: To what extent does the language background of test takers affect their performance in content-based assessment delivered in English? To what extent do test takers' English language abilities (especially reading and writing skills) affect their performance in such assessments? To what extent is test takers' performance affected by language complexity and the linguistic demands of test items? What are the interactive effects on test performance of item features with test takers' language abilities? How might test accommodations (e.g. modifying linguistic demands of test items) affect performance in content-based assessment? From this list of key questions, test accommodations, whilst fraught with problems, seem to be able to provide the most

direct and immediate policy remedy to address the complexity of issues surrounding 'language in examinations'. Educators have been seeking, particularly since the implementation of the *No Child Left Behind Act of 2001* in the USA, opportunities for providing linguistically disadvantaged students with a variety of test accommodations in formal large-scale assessments of content-knowledge, in order to 'level the playing field' (Thurlow and Bolt 2001) with mainstream students (Wolf, Kao, Herman, Bachman, Bailey, Bachman et al 2008), as explained below.

Test accommodations

The concept of test accommodation originates from the assessment of students within the domain of special needs. Thurlow and Bolt (2001:1) define accommodations as:

> . . . changes in assessment materials or procedures that address aspects of students' disabilities that may interfere with the demonstration of their knowledge and skills on standardized tests. Accommodations attempt to eliminate barriers to meaningful testing, thereby allowing for the participation of students with disabilities.

In the context of assessing learners with EAL, Koenig and Bachman (2004:1) define accommodations as actions 'taken in response to a determination that an individual's disability or level of English language development requires a departure from established testing protocol'. Similarly, Butler and Stevens (1997:5) define accommodations for EAL learners on large-scale content assessments as:

> . . . support provided [to] students for a given testing event, either through the modification of the test itself or through the modification of the testing procedure, to help students access the content in English and better demonstrate what they know.

In theory, Wolf et al (2008) suggest three criteria for the use of accommodations: effectiveness, validity and feasibility. In other words, an accommodation should be effective in raising the performance of those who should receive the accommodation, but should not change the nature of the test task, nor should it provide those who receive it unfair advantages over their peers. Further, those who need the accommodation should benefit but not those who do not (Sireci, Li and Scarpati 2003). The ultimate purpose of providing accommodations is, therefore, to produce valid assessment outcomes, in addition to levelling the playing field for those learners considered at risk or 'marginalised' in some way. The biggest challenge in implementing test accommodations is, however, how to achieve this kind of sensitive balance.

In practice, test accommodations for EAL learners usually refer to changes to the test itself and/or the test procedure (Butler and Stevens 1997:6, see Table 1 below). Types of test accommodations have also been classified differently (e.g. by the team at the National Center of Educational Outcomes at the University of Minnesota, USA) as 'presentation' (e.g. linguistic modification of test items, dual language), 'timing' (extra time, multiple sessions), 'response' (booklet vs answer sheet) and 'setting' (e.g. separate room).

Table 1 Exemplar Accommodation types

Modifications of the test	Modifications of the test procedure
• assessment in the native language	• extra assessment time
• test change in vocabulary	• breaks during testing
• modification of linguistic complexity	• administration in several sessions
• addition of visual supports	• oral directions in the native language
• use of glossaries in native language	• small-group administration
• use of English glossary	• separate-room administration
• linguistic modification of test directions	• use of dictionaries
• additional example items/tasks	• reading aloud of questions in English
	• answers written directly in test booklet
	• directions read aloud or explained

Source: Butler and Stevens (1997:6). Used with permission.

There is now a growing body of empirical research on test accommodations, especially from the USA (e.g. Abedi and Hejri 2004, Abedi, Hofstetter and Lord 2004, Abedi, Lord, Hofstetter and Baker 2000, Albus, Thurlow, Liu and Bielinski 2005, Bielinski, Thurlow, Ysseldyke, Freidebach and Freidebach 2001, Bolt and Ysseldyke 2006, Butler and Stevens 1997, Duncan, Parent, Chen, Ferrara, Johnson, Oppler et al 2005, Elbaum 2007, Hofstetter 2003, Koenig and Bachman 2004, Kopriva, Emick, Hipolito-Delgado and Cameron 2007, Stansfield 2002). Research on test accommodations in SSA contexts is rare.[3]

A wide range of different test accommodations initiatives (Abedi 2008, Francis, Rivera, Lesaux, Kieffer and Rivera 2006) have been implemented. Here we focus on linguistic accommodations, i.e. those modifications directly related to the 'language' of the test rather than testing procedures. We also limit our review to the effects of the most widely used linguistic accommodations for non-disabled students.[4] Several extensive reviews of accommodation studies in the USA context have already been published (e.g. Abedi, Hofstetter and Lord 2004, Francis, Rivera, Lesaux, Kieffer and Rivera 2006, Sireci, Li and Scarpati 2003, Wolf et al 2008).[5] In this paper, we focus on the effects of linguistic modification of items, dual-language or side-by-side bilingual test versions, and tests conducted in the L1 on learner performance, drawing on the literature from both developed and developing countries (in particular SSA).

Linguistic modifications refer to the changes made to test items to reduce

the construct-irrelevant linguistic complexity and demands, as one of the ways of reducing bias of a linguistic nature, so as to facilitate test takers who would be otherwise disadvantaged because of their low language proficiency (Sireci and Allalouf 2003, Uiterwijk and Vallen 2005), while maintaining the construct of the test. Linguistic modifications may include using familiar words, shorter and simple sentences, and increasing the readability of test items (see Abedi, Lord and Plummer 1997). Studies using linguistic modifications for learners with EAL in content-based assessment have drawn conceptually on findings of psychological studies where a change in the wording or structure of a test item (e.g. mathematics word problems) has been observed to affect students' performance (usually L1 learners; e.g. Aiken 1971, 1972, Cummins, Kintsch, Reusser and Weimer 1988, De Corte, Verschaffel and De Win 1985, Larsen, Parker and Trenholme 1978, Lepik 1990, Wheeler and McNutt 1983, to name just a few). However, in practice, the effects of using linguistic modifications on test performance of EAL learners appear much more complex. For example, Abedi, Lord and Plummer (1997) found that students at different ability levels in maths benefited differentially from linguistic modifications: with students in low and average maths classes scoring higher, albeit not significantly different, in the linguistically modified test. Abedi and Lord (2001) found that modifying linguistic structures in maths word problems (especially in relation to the frequency of vocabulary and the voice of verb constructions) affected students' performance; and that these effects were small but significant for low-performing EAL students. In other studies by Abedi and colleagues (Abedi, Hofstetter, Baker and Lord 2001, Abedi, Lord, Hofstetter and Baker 2000, Abedi, Lord and Hofstetter 1998), linguistic modifications (e.g. using shorter problem statements) in maths test items were found to be generally capable of improving EAL students' test scores, albeit not significantly. In the Abedi, Courtney, Leon, Kao and Azzam (2006) study, linguistic modifications were not found to impact on students' maths performance. EAL students' test scores were also improved in linguistically modified science items (e.g. Abedi, Courtney and Leon 2003). However, these effects varied for different grades: performance differences were seen for Grade 8 students, but not for Grade 4 students (see also Abedi and Lord 2001). Abedi, Courtney, Mirocha, Leon and Goldberg (2005) observed that linguistic modifications in science test items were more effective for Grade 8 than Grade 4 students in narrowing the performance gap between EAL and non-EAL students (see also Abedi, Courtney and Leon 2003). Yet, Rivera and Stansfield (2004) found that linguistic modifications (or in their terms, linguistic simplification) did not facilitate significant improvement of either Grade 4 or Grade 6 EAL students' performance in science tests. However, it should be noted that this study had a very small sample of EAL students for statistically meaningful comparisons. In African contexts, Prophet and Badede (2006) made changes to test items, such as the

readability (see also Shorrocks-Taylor and Hargreaves 1999, 2000 in the UK context) and length of questions, and changes of some words, tense or structure of questions and found that the effects on student performance were statistically significant for the readability of MCQs: improving readability enhanced student performance. Shortening the length of the stem, changing to past tense, removing unnecessary information, replacement of unfamiliar or vague words all made some difference, with gains noted for both boys and girls. As Abedi, Hofstetter and Lord (2004:11) rightly pointed out, the effects of linguistic modification vary and 'may have depended in part on the nature and extent of the modifications made'. Additionally, the differential effects of even identical linguistic modifications in different subject areas at different educational levels in content knowledge pose further constraints on the usefulness and the generalisability of such findings.

Other types of linguistic accommodation involve translation of test items into the learners' L1, or dual-language or side by side bilingual test versions as an alternative to test translation (Solano-Flores, Trumbull and Nelson-Barber 2002). Duncan et al (2005) and Abedi, Courtney, Leon, Kao and Azzam (2006) both found dual-language maths tests did not affect student performance, although they were preferred by the test takers (Duncan et al 2005). In the African context, Eisemon, Schwille, Prouty, Ukobizoba, Kana and Manirabona (1993) in Burundi investigated how language affected the measurement of literacy, mathematics and science achievement of sixth-grade students in predominantly rural areas of Burundi. A multi-stage stratified cluster sample was drawn by probability methods from 21 cantons of the country. Tests were developed to assess student achievement in the domains of reading comprehension, written composition in narratives, mathematics and science (including elements of agriculture and health as well). The tests were initially developed in English and translated into French and then from French to Kirundi. Five versions of the tests were administered to subsamples as follows: French comprehension and composition (with standard French) to ⅛ of the students; French comprehension and composition (simplified, colloquial French): ⅛; Kirundi comprehension and composition: ¼; French mathematics, science and agriculture: ¼; Kirundi mathematics, science and agriculture: ¼. Students were given two hours to finish one of the five versions. The language of assessment, French or Kirundi, was observed to profoundly influence the measurement of achievement in most of the subjects tested. In the Kirundi versions of comprehension, composition and science/agriculture, students achieved significantly higher scores, with the greatest difference observed for the science/agriculture test. In addition, the performance of the most able students was the most affected by being measured through French. Only in mathematics were the results from the French and Kirundi versions nearly identical.

Leaving aside the potentially different impact of the use of first language

and dual-language test versions, as reviewed above, there are serious theoretical, methodological and practical limitations in translating tests, in particular, the comparability in the psychometric properties between the different language forms of a test (Allalouf, Hambleton and Sireci 1999, Ercikan and Koh 2005, Sireci and Allalouf 2003, Sireci and Khaliq 2002).

In summary, the review of the literature on the effects of accommodations, in particular linguistic modifications of test items, on test performance not only reveals the significant conceptual and methodological challenges of applying accommodations in a reliable, valid and feasible manner, it also demonstrates the inconsistency in the findings and, as Wolf et al (2008:41) rightly conclude, the research 'provides little evidence to assure valid procedures for applying accommodations'. This is probably due to the complex nature and the interactions of accommodations with numerous other factors in educational assessment such as subjects (e.g. science, mathematics, social science) which have different linguistic demands, student characteristics (e.g. language proficiency, subject knowledge) and the ways in which the accommodations are implemented across different studies and educational contexts (e.g. whether in the USA and UK or SSA contexts). Two further notable limitations of the accommodations studies reviewed are: (a) they have focused mainly on the product of learning in terms of test performance and score comparability. Studies on the internal cognitive processes of test takers when provided with dual-language tests are much needed, for instance, how their L1 and L2 resources are utilized (e.g. Cohen 1994); (b) they have failed to look beyond accommodations. The evidence for poor examination performance for L2 learners requires that we look beyond the potential impact of accommodations in reducing discrimination and unfairness in high-stakes examinations, and indeed beyond examinations. Assessment accommodations may be able at best to provide a quick fix, but accommodations alone are not able to address the fundamental challenges for students studying through an unfamiliar language and the reasons behind their underperformance or, indeed, whether the issues of underachievement are a reflection of the school system as a whole. It is also crucial to analyse the possible impact of instruction-embedded factors and the ways in which learners' examination performances may be shaped by the nature of language use in teaching and learning in order to gain better understandings of the interrelationships between the language of assessment and the language of instruction. This is the focus of the next section.

Language in classroom learning

Below, we draw on findings from empirical studies referenced to three aspects of classroom learning and language use: code switching/mixing and discourse features of teacher talk, classroom pedagogy and learner classroom language. We do this so as to investigate the nature of classroom interaction

and how this might impact on student learning and, ultimately, their ability to show their achievement (language and content) through formal examining mechanisms. The empirical studies that form the basis of this review have gathered data in both subject (e.g. science and mathematics) and language learning classrooms, the majority of which have been undertaken in African schools.

Teacher language

Code switching and mixing

Code switching and mixing has been the focus of research in multilingual classrooms for some considerable time. Early identification of this phenomenon in the SSA context was largely brought to prominence by the research of Rubagumya (1991, 1993, 1994). Research to date has shown that teachers are motivated in their use of L1 and L2 in their teaching by a range of reasons. For example, Probyn (2005, 2006) investigated the classroom practices of six Grade 8 (first year of secondary school) science teachers teaching through the medium of English. In her studies, she found that code switching was associated with those occasions when teachers were engaged in explaining new concepts, clarifying statements or questions, emphasising particular points, and making connections between learners' own contexts and experience. These findings are similar to those reported by Brock-Utne (2002) in her analysis of Mwinsheikhe (2001), where 82% of teachers who admitted using L1 in their lessons reported doing so in order to clarify difficult and/or key concepts of the lessons. Additional reasons are provided by Cleghorn and Rollnick (2002) based on empirical data from a number of studies in African classrooms that evidenced code-switching as the means to clarify linguistically based confusion, render the culturally unfamiliar familiar, make the implicit explicit, provide English vocabulary needed for examination purposes, provide contextualisation cues, and raise learners' metalinguistic awareness. These examples, thus, point to the potential cognitive advantages of the use of the L1 in learning. Other uses of code switching have been linked with classroom management and discipline (Probyn 2005) and with providing instructions for practical work and assignments Brock-Utne (2002). Translation has also been identified separately as a teacher's coping strategy by Brock-Utne and Holmarsdottir (2004). Some research studies have also associated code switching with an affective dimension, in addition to its potential impact in learning, such as creating a good atmosphere (Brock-Utne 2002), and maintaining learners' attention (Probyn 2005).

Some of these studies, as well as those that have investigated the effects of teaching and learning in the mother tongue, have identified benefits from using the L1 for which there is scant evidence in the L2 classroom. These have been summarised by Heugh (2000) and include enhanced classroom

participation, positive affect and increased self-esteem (Dalby 1985, Dutcher 1995, Richardson 2001), increased parent participation (Cummins 2000, Dutcher 1995) and participation of girls (Benson 2002, Hovens 2002), all of which have significant potential impact in relation to equality of access to knowledge and educational opportunity, especially for girls.

Whilst the above examples identify how code switching and mixing may be used to enhance classroom learning, there is also evidence that suggests that teachers code switch in order to accommodate the ability of their learners in the medium of instruction (e.g. Arthur and Martin 2006, Ndayipfukamiye 1994, Probyn 2005, Setati, Adler, Reed and Bapoo 2002), as well to compensate for their own weak command of the L2 (Brock-Utne 2005). Others have found that the use of code switching may vary according to teacher expertise. Brock-Utne (2005), for example, reported higher use of this strategy by teachers who were not language teachers, a finding that corresponds to those of Setati et al (Setati et al 2002), in their longitudinal case study research of language practices in maths, science and English, who noted the prevalence of code switching by both teachers and learners in the maths and science classes.

We take from the above that code switching and the use of the L1 may be used to facilitate the teaching and learning process as well as to create a conducive learning environment. The research also tells us that use of the L1 may operate in default mode on account of teacher and/or learner weaknesses in the target language. Finally, although not reported here in any detail, it becomes clear from the literature that the extent of code switching and mixing in African classrooms, in both language and content classes, may be extremely high (e.g. Brock-Utne and Holmarsdottir 2004, Heugh 2000, Probyn 2005), a feature that may impact on the extent to which learners are prepared to engage with the demands of examinations (see discussion below).

Classroom discourse

From the research studies reviewed, one ubiquitous finding is that much teacher L2 talk is very restricted, although Probyn (2005) – as an exception – reported rich linguistic input in some of the science classes observed. Hornberger and Chick (2001) write about 'safe talk', that has emerged as common practice in African classrooms, largely a reflection of the teachers' own limited proficiency in the language of instruction. According to Rubagumya (2003), this is characterised by the way teachers encourage chorusing from their learners, and the repeating of phrases or words (after the teacher) and copying notes from the blackboard. Another example (Probyn 2005) occurs when teachers read aloud to their learners from notes or a textbook. Cleghorn and Rollnick (2002) have also identified a facet of washback from examinations on teaching where past examination papers take the

place of lesson plans, teacher creativity, and curriculum innovation. Because the examinations are often multiple choice in format or require short, fixed answers, teacher instruction tends to emphasise rote question-and-answer routines requiring single-word answers in order to prepare learners to recognise the key (English) words on the exam.

Teachers' poor command of the L2 as well as the limited amount of L2 actually used in the classroom has been raised by numerous researchers (e.g. Qorro 2002, cited in Brock-Utne 2005, Rugemalira 2005), which, in turn, may prevent teachers from articulating the subject matter clearly (e.g. Cleghorn and Rollnick 2002, Probyn 2006, Roy-Campbell 1995). Other studies have identified classroom talk as being almost exclusively by teachers (e.g. Bunyi 2005, Alidou and Brock-Utne 2006) and how this talk takes place from the front of the classroom (Arthur and Martin 2006). This links with the data from Kapp (2004) where both teachers in his/her study placed themselves at the centre of classroom activity, asking questions and vetting answers.

This paucity, not only in amount but also range of teachers' use of the L2, has been discussed by Prophet and Badede (2006:4) who observe how 'the teachers themselves are teaching as 2nd or 3rd language speakers and thus their linguistic reservoir is shallow'. The effects of this on learners have been reported by Arthur (1994) and Alidou and Brock-Utne (2006) who found that teachers adopt practices which reduce the cognitive demands of lessons, i.e. evidence of content 'watering down'. In one of the few studies that analyses the nature of teacher questioning, Probyn (2005) reported a greater incidence of higher order questions in one out of the six classes observed. In the others, however, the dominant types of questions were those that encouraged student responses requiring recall and review of prior knowledge and work.

There appears little evidence in the studies of L2 classroom discourse reviewed that provides a strong scaffold for either content or language learning. One example that exemplifies a move from informal spoken and exploratory talk as the basis for formal discourse-specific writing is reported by Setati et al (2002), although they note this is complicated by the fact that learners' exploratory talk may be in their L1 rather than the language of instruction. In terms of feedback to learners, Kapp (2004) observed that both teachers of EAL in his study seldom nominated individuals or asked them to restructure or clarify when there was a communication breakdown. Further, Probyn (2005) found that when a teacher restricted her use of the L2 to direct reading from the textbook, it was assumed that the learners should understand the content and the teacher did not provide any strategies as to how her learners could move from her oral Xhosa presentation and their need to read and write and be assessed in English which, it was asserted, was the responsibility of the English rather than the subject teacher. Such findings, however, contrast with the case of bilingual teachers (e.g. Benson 2002), who were able to get more immediate and comprehensible feedback about what students

know and what they are learning and, as a consequence, can make more realistic evaluations of their students' performance. She also found that these teachers, i.e. those who transitioned between the L1 and L2, were less likely to judge students or to generalise – for example that all girls are less able – because they have more evidence of what their students can and cannot do.

What emerges from our review of teacher discourse is that many teachers in SSA classrooms struggle in their use of the L2 and, as a consequence, are unable to provide rich linguistic environments for their L2 learners through which they are able to develop both their L2 language abilities and their subject knowledge and skills which are assessed in formal examinations through the L2. These findings of teacher language use would appear to differ in several respects from studies that have investigated dual language and bilingual classroom contexts (e.g. Dalby 1985, Dutcher 1995, Richardson 2001, Ouane 2003) in which, as evidenced by Hovens (2002), with reference to case studies in Niger and Guinea-Bissau that bilingual classrooms were more stimulating, interactive, and relaxed, as well as bringing considerable cognitive advantage.

Classroom pedagogy

Given the findings reported above of the problems faced by teachers in using the L2, it is unsurprising to find that instances of pedagogy that promotes sustained interaction in learning are minimal. For example, several studies have reported an absence of group work (e.g. Arthur 1994, Arthur and Martin 2006, Setati et al 2002). By way of an exception, over the three years of their research, Setati et al (2002) and colleagues observed the increasing use of group work in most of their classes and evidence of 'learning talk'. However, this increase in oral work was accompanied by limited writing of extended texts in English with written work restricted to exercises, and few opportunities for learners to use and develop spoken and written English, as they could revert to the L1. Kapp (2004), too, observed the lack of follow-up with writing tasks, and suggested that the teachers believed that writing ability would flow naturally from oral competence. These findings resonate with several other studies (e.g. Probyn 2006) that report restricted opportunities for learners in reading, writing and speaking, especially activities that would develop exploratory and explanatory skills, and allow learners to take risks using the L2 (e.g. Setati et al 2002, Probyn 2006, Alidou and Brock-Utne 2006). A further finding with implications for curriculum access and support is the tendency for teachers to interact with the same small numbers of learners (e.g. Arthur 1994, Myers-Scotton 1993, Kapp 2004).

Lack of resources to support learning more generally and the development of reading skills in particular is undoubtedly linked to their very limited availability to teachers (e.g. Arthur 1994, Mbise 1994, Probyn 2006). Probyn

(2005) also explained that where textbooks are available they were difficult to understand, particularly science textbooks (Ogunniyi 2005). Further, given the extremely short supply of materials for learners, Cleghorn and Rollnick (2002) have observed how teachers may use the textbook to create oversimplified if not misleading worksheets.

In summary, a restricted range of pedagogic activities to engage learners in subject and language learning is reported within environments with very limited resources available to both teachers and students to support learning.

Learning through an L2

A constant refrain and finding from studies spanning several decades points to the fact that learners in African L2-medium classrooms often cannot speak the L2 well enough to use it as a medium for learning (e.g. Criper and Dodd 1984, Macdonald 1990, Dutcher 1995, Williams and Cooke 2002, Alidou and Brock-Utne 2006). The early research of MacDonald in South Africa (1990) pointed to learners' low writing and reading abilities, with more recent evidence for the latter provided by Broom (2004) from South Africa and from the systematic series of reading research studies in Zambian primary schools (Williams 1996, 2006).

In his investigation of ESL literacy practices, Kapp (2004) found evidence of learners acquiring conversational fluency and confidence in English (but see Cummins 1984, Thomas and Coliier 1997). However, the literacy they learned was functional and situation-specific with students having little or no opportunity to reflect and process at a cognitively demanding level in English in either oral or written form. Difficulties with written genres at school were reported by Kapp (2004), in particular the difficulties that learners had with constructing a written argument, tending to summarise, and to 'read on the line', rather than analyse the language of the literary texts. Neither is there evidence that they have the conceptual framework, the metalinguistic tools, or the vocabulary required for the task. Further, Probyn's research (2005, 2006) identified ways in which teachers' questioning and their support strategies impacted on the nature and patterns of student response in terms of amount and quality of learner responses.

In terms of oral interactions, research highlights how students, in both content and language lessons, immediately switched to L1 in group discussions (e.g. Brock-Utne 2002). She further reported that 63% of teachers in Mwinsheikhe's (2002) study allowed students to use L1 in their lessons and that they normally used L1 during group work. Interestingly, the students admitted that they followed lessons better if Kiswahili was used.

In addition to limitations at the level of L2 language skills, several studies have identified L2 system related difficulties faced by learners, especially in

science subjects and maths. For example, Prophet and Badede (2006) reported research that focused on the use of non-technical words in science (e.g. citing Cassels and Johnstone 1983, 1985). Fang (2006) identified the following as potentially problematic categories for learners: use of technical vocabulary, ordinary words with non-vernacular meaning or usages, prepositions, conjunctions and pronouns, ellipsis, subordinate clauses, with-prepositional phrases, abstract and lengthy nouns, complex sentences, interruption constructions, and passive voice (see also Hazel, Logan and Gallagher 1997).

The effects, then, of learning through an unfamiliar language of instruction and the evidence of teacher-centred interaction in which learners are neither stretched nor scaffolded in their attempts to use the L2 creatively in responses (oral or written) much beyond the single sentence has the effect that the learners in class are predominantly silent, any participation tending to take the form of choral repetition and recall, i.e. the 'safe talk' phenomenon (Hornberger and Chick 2001); see also Alidou 1997, 2003, Brock-Utne et al 2005, Brock-Utne 2005, Hovens 2002, Rubagumya 2003). Commenting on the limited participation of students in class, even when given the opportunities to respond, teachers explained this in terms of students not knowing how to answer in English even though they could understand what was asked of them (Roy-Campbell 1995).

The findings presented above in relation to classroom language use and opportunities for using the L2 in subject and language learning carry significant implications for the way in which learners are able to tackle school formal examinations, as discussed below.

Some conclusions

The impact of these findings from the African classroom on the ability of learners to perform well in formal written examinations is not hard to explain: students are inadequately prepared for examinations. In class, students are not exposed to rich language models. Much of the lesson may be in the L1 and where this is not the case, they have little opportunity to exchange meanings in the L2 in science and mathematics learning, with their outputs largely restricted to single sentence or word answers. They do not have rich exposure to the language of the examinations, as they are not stretched to develop their skills of reasoning, explanation or justification that is required in the development of conceptual understandings as well as in the construction of the examination responses. In other words, students may lack exposure to and use of the variety of genre for learning, as opposed to teacher questioning that 'tests' recall and memorisation of information. They may lack, as an example, problem solving activities where students are presented with a problem in maths or science and asked to solve it, thus requiring them to explain (e.g. sequential or causal explanations), argue (e.g. Erduran and

Jiménez-Aleixandre 2007), or organise information through describing, reporting, discussing or evaluation.

The teachers' own linguistic skills may be severely limited such that learners are not scaffolded in their attempts to develop their content knowledge and language skills and one of the results may be that the subject content itself is simplified and watered down. The effects of 'safe talk', and a linguistically and interaction-reduced classroom environment, as evidenced in the empirical findings reported above, have profound effects on learner participation with few opportunities to use language productively for learning through the L2 as in group or class discussions, pair work, or reporting back.

The social consequences for these learners are considerable: they will be forced to leave school, on the basis of poor examination performance labelled low achievers whereas, in reality, it is many of the factors within the teaching and learning context that inhibit the extent to which they may acquire subject knowledge and develop the skills which will enable them to reach their potential.

There is also a range of non-linguistic factors (not discussed in this paper) that impact on students' abilities to achieve in formal examinations such as the impoverished and illiterate backgrounds from which many learners come, the extent to which teachers are transferred from one school to another, teacher absenteeism, and limited opportunities for teachers to develop their professional skills (Yu and Thomas 2008). Notwithstanding these inhibiting factors in the provision of education quality, this paper has highlighted a number of language parameters that impact on the extent to which school age learners are able to demonstrate their school achievements through the medium of an unfamiliar language.

The case for inequities in educational access and assessment in SSA is strong. In terms of classroom engagement, there is evidence of educational advantage in classes where a familiar or dual languages are used in instruction: bilingual classrooms have been observed to be more stimulating and interactive, with significant gains observed by those participating in bilingual programmes particularly rural children and girls. Hovens (2002), for example, reported that in 64 bilingual classes in a study in Mozambique, learners had the courage to ask questions or even correct the teacher (which was never observed in traditional classrooms). There are benefits in terms of learner self-esteem issues related to the fact that children are allowed to express their full range of knowledge and experience in a language in which they are competent.

By way of contrast, as evidenced through our review of primarily empirical studies, educational disadvantage prevails in the majority of classes conducted through the medium of a language that is not their mother tongue, with evidence showing that learners face considerable cognitive and linguistic challenges in acquiring conceptual understandings across the curriculum,

which impact on their engagement and participation in class. Heugh (2000: 12) summarises the situation:

> There is in fact a huge body of research which has been conducted in South Africa and which points conclusively to the disastrous effects of attempting to teach mainly through English when the conditions do not and cannot make this possible.

However, the fact remains that in many SSA countries national examinations at secondary level are delivered through the medium of English and this will remain to be the case for some considerable time to come. The consequences are very high for each and every child. It therefore becomes imperative that, at the very least, some attempt is made to address the current instructional limitations identified above as well as to research further ways in which school age learners may be enabled to show what they know, i.e. their content knowledge and skills. Ensuring quality education processes, of which examining (written examinations as well as teacher assessment) forms a significant part, and facilitating continued access to education opportunities beyond basic education both address some of the goals of the EFA. They also have the potential to respond to issues of disadvantage that have direct consequences on individual learners in instructional contexts where an unfamiliar language is used both for instructing and examining. The stakes and the consequences are both very high!

Notes

1. EAL (English as an Additional Language), ESL (English as a Second Language) and ELL (English Language Learners) are used to refer to learners who are studying in school through a language that is not their first/heritage language. In this chapter, we use EAL to refer to these students.
2. It is to be noted, in contrast, that Heugh (2000) showed that in the Western Cape where 80% of learners wrote their 1999 matriculation examination in their mother tongue, there was a 79% pass rate.
3. Student Performance in National Examinations: the dynamics of language (SPINE) focuses on the use of both Kiswahili and English in the scholastic achievement of learners at the end of Basic Education in Zanzibar. One in a series of studies focuses on the effects of linguistic accommodations within the context of the national Form II examinations (i.e. at the end of Basic Education). Funded by ESRC/DfID (2007–2010), www.bristol.ac.uk/spine
4. For test accommodations for students with physical and cognitive disabilities, see the reviews conducted at the National Center on Educational Outcomes at the University of Minnesota (e.g. Johnstone, Altman, Thurlow and Thompson 2006; Thompson, Blount and Thurlow 2002; Zenisky and Sireci 2007) and papers in the special issue (vol. 31, no.1) of *Assessment for Effective Intervention*.

5. Sireci et al (2003) mainly reviewed studies on test accommodations for students with special needs, but also some studies on accommodations for ELLs in the USA.

References

Abedi, J (2002) Standardized achievement tests and English language learners: psychometrics issues, *Educational Assessment* 8 (3), 231–257.

Abedi, J (2008) Utilizing accommodations in assessment, in Shohamy, E and Hornberger, N H (Eds) *Encyclopaedia of Language and Education, Vol 7: Language testing and assessment*, 331–347.

Abedi, J and Gándara, P (2006) Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy, *Educational Measurement: Issues and Practice* 25 (4), 36–46.

Abedi, J and Hejri, F (2004) Accommodations for students with limited English proficiency in the national assessment of educational progress, *Applied Measurement in Education* 17 (4), 371–392.

Abedi, J and Lord, C (2001) The language factor in mathematics tests, *Applied Measurement in Education* 14 (3), 219–234.

Abedi, J, Courtney, M and Leon, S (2003) *Effectiveness and validity of accommodations for English language learners in large-scale assessments (CSE technical report no. 608)*, Los Angeles: University of California Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J, Courtney, M, Leon, S, Kao, J and Azzam, T (2006) *English language learners and math achievement: A study of opportunity to learn and language accommodation. CRESST technical report 702*, Los Angeles: University of California, Los Angeles.

Abedi, J, Courtney, M, Mirocha, J, Leon, S and Goldberg, J (2005) *Language accommodations for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification (CSE technical report no. 666)*, Los Angeles: University of California Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J, Hofstetter, C H, Baker, E and Lord, C (2001) *NAEP math performance and test accommodations: Interactions with student language background (CRESST tech. Rep. No. 536)*, Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J, Hofstetter, C H and Lord, C (2004) Assessment accommodations for English language learners: Implications for policy-based empirical research, *Review of Educational Research* 74 (1), 1–28.

Abedi, J, Lord, C, Hofstetter, C and Baker, E (2000) Impact of accommodation strategies on English language learners' test performance, *Educational Measurement: Issues and Practice* 19 (3), 16–26.

Abedi, J, Lord, C and Hofstetter, C H (1998) *Impact of selected background variables on students' NAEP math performance (CSE technical report no. 478),* Los Angeles: University of California, Center for the Study of Evaluation/ National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J, Lord, C and Plummer, J (1997) *Final report of language background as a variable in NAEP mathematics (CSE tech. Report no. 429)*, Los Angeles, CA:

University of California Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.
Aiken, L R Jr (1971) Verbal factors and mathematics learning: A review of research, *Journal for Research in Mathematics Education* 2 (4), 304–313.
Aiken, L R Jr (1972) Language factors in learning mathematics, *Review of Educational Research* 42 (3), 359–385.
Albus, D, Thurlow, M, Liu, K and Bielinski, J (2005) Reading test performance of English-language learners using an English dictionary, *Journal of Educational Research* 98 (4), 245–254.
Alidou, H (1997) *Education language policy and the bilingual education: the impact of French language policy in primary education in Niger*, unpublished PhD Thesis dissertation, University of Illinois Urbana-Champaign.
Alidou, H (2003) Language policies and language education in francophone Africa: a critique and a call to action, in Makoni, S, Smitherman, G, Ball, A F and Spears, A K (Eds) *Black Linguistics: Language, Society, and Politics in Africa and the Americas*, London/NY: Routledge.
Alidou, H and Brock-Utne, B (2006) Experience I – teaching practices – teaching in a familiar Language, in Alidou, H et al *Optimizing Learning and Education in Africa – the Language Factor*: *A Stock-taking Research on Mother Tongue and Bilingual education in Sub-Saharan Africa,* Paris: ADEA.
Allalouf, A, Hambleton, R K and Sireci, S G (1999) Identifying the causes of DIF in translated verbal items, *Journal of Educational Measurement* 36 (3), 185–198.
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999) *Standards for educational and psychological testing*, Washington, DC: American Educational Research Association.
Arthur, J (1994) English in Botswana primary classrooms: functions and constraints, in Rubagumya, C (Ed.) *Teaching and Researching Language in African Classrooms*, Clevedon: Multilingual Matters Ltd, 63–78.
Arthur, J and Martin, P (2006) Accomplishing lessons in postcolonial classrooms: comparative perspectives from Botswana and Brunei Darussalam, *Comparative Education* 42 (2), 177–202.
Bailey, A L and Butler, F A (2004) Ethical considerations in the assessment of the language and content knowledge of U.S. School-age English learners, *Language Assessment Quarterly* 1 (2 & 3), 177–193.
Benson, C J (2002) Real and potential benefits of bilingual programmes in developing countries, *International Journal of Bilingual Education and Bilingualism* 5 (6), 303–317.
Bielinski, J, Thurlow, M, Ysseldyke, J E, Freidebach, J and Freidebach, M (2001) *Read-aloud accommodation: Effects on multiple-choice reading and math items (technical report 31)*, Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.
Bolt, S E and Ysseldyke, J E (2006) Comparing DIF across math and reading/language arts tests for students receiving a read-aloud accommodation, *Applied Measurement in Education* 19 (4), 329–355.
Brock-Utne, B (2001) Education for all – in whose language? *Oxford Review Of Education* 27 (1), 115–134.
Brock-Utne, B (2002, 7–9 January) *The most recent developments concerning the debate on language of instruction in Tanzania*, paper presented at the NETREED Conference, University of Oslo.
Brock-Utne, B (2005) Language-in-education policies and practices in Africa

with a special focus on Tanzania and south Africa – insights from research in progress, in Lin, A M Y and Martin, P W (Eds) *Decolonisation, globalisation: Language-in-education policy and practice,* Multilingual Matters, Clevedon: England,173–193.
Brock-Utne, B and Holmarsdottir, H B (2004) Language policies and practices in Tanzania and south Africa: Problems and challenges, *International Journal of Educational Development* 24 (1), 67–83.
Brock-Utne, B, Desai, Z and Qorro, M (Eds) (2005) *LOITASA Research in progress*, Dar es Salaam: KAD Associates.
Broom, Y (2004) Reading English in multilingual south African primary schools, *International Journal of Bilingual Education and Bilingualism* 7 (6), 506–528.
Bunyi, G W (2005) Language classroom practices in Kenya, in Lin A M Y and Martin P W (Eds) *Decolonisation, Globalisation: Language-in-Education Policy and Practice*, Multilingual Matters, Clevedon: England, 131–152.
Butler, F A and Stevens, R (1997) *Accommodation strategies for English language learners on large scale assessments: Student characteristics and other considerations. (CSE tech. Report no. 448)*, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
Cassels, J, and Johnstone, A H (1983) Meaning of words and the teaching of chemistry, *Education in Chemistry*, 20 (1), 10–11.
Cassels, J R T and Johnstone, A H (1985) *Words that matter in science*, London: Royal Society of Chemistry.
Clayton, T (1998) Explanations for the use of languages of wider communication in education in developing countries, *International Journal of Educational Development* 18 (2), 145–157.
Cleghorn, A and Rollnick, M (2002) The role of English in individual and societal development: A view from African classrooms, *TESOL Quarterly* 36 (3), 347–372.
Cohen, A D (1994) The language used to perform cognitive operations during full-immersion maths tasks, *Language Testing* 11 (2), 171–196.
Criper, C and Dodd, W (1984) *Report on the teaching of English language and its use as a medium of instruction in Tanzania*, Dar es Salaam: The British Council.
Cummins, D D, Kintsch, W, Reusser, K and Weimer, R (1988) The role of understanding in solving word problems, *Cognitive Psychology* 20 (4), 405–438.
Cummins, J (1984) *Bilingualism and special education: Issues in assessment and pedagogy*, Clevedon: Multilingual Matters.
Cummins, J (2000) *Language, power and pedagogy: Bilingual children in the crossfire*, Clevedon: Multilingual Matters.
Dalby, D (1985) *The educational use of African languages in sub-Saharan Africa: The state of the art*, Paris: UNESCO.
De Corte, E, Verschaffel, L and De Win, L (1985) Influence of rewording verbal problems on children's problem representations and solutions, *Journal of Educational Psychology* 77 (4), 460–470.
Duncan, T G, Parent, L d R, Chen, W-H, Ferrara, S, Johnson, E, Oppler, S et al (2005) Study of a dual-language test booklet in eighth-grade mathematics, *Applied Measurement in Education* 18 (2), 129–161.
Dutcher, N (1995) *Expanding educational opportunities in linguistically diverse societies*, Washington D.C.: Center for Applied Linguistics.

Eisemon, T O, Schwille, J, Prouty, R, Ukobizoba, F, Kana, D and Manirabona, G (1993) Providing quality education when resources are scarce: Strategies for increasing primary school effectiveness in Burundi, in Levin, H M and Lockheed, M E (Eds) *Effective schools in developing countries,* London: The Falmer Press (with the World Bank),130–157.

Elbaum, B (2007) Effects of an oral testing accommodation on the mathematics performance of secondary students with and without learning disabilities, *Journal of Special Education* 40 (4), 218–229.

Ercikan, K and Koh, K (2005) Examining the construct comparability of the English and French versions of TIMSS, *International Journal of Testing* 5 (1), 23–35.

Erduran, S and Jiménez-Aleixandre, M P (2007) *Argumentation in science education: Perspectives from classroom-based research*: Springer.

Fang, Z (2006) The language demands of science reading in middle school, *International Journal of Science Education* 28 (5), 491–520.

Francis, D J, Rivera, M, Lesaux, N, Kieffer, M and Rivera, H (2006) *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessment*, Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Greaney, V and Kellaghan, T (1995) *Equity issues in public examinations in developing countries,* Washington, DC: The World Bank.

Hargreaves, E (1997) Mathematics assessment for children with English as an additional language, *Assessment in Education: Principles, Policy & Practice* 4 (3), 401–412.

Hazel, E, Logan, P and Gallagher, P (1997) Equitable assessment of students in physics: Importance of gender and language background, *International Journal of Science Education* 19 (4), 381–392.

Heugh, K (2000) *The case against bilingual and multilingual education in south Africa*, Cape Town: University of Cape Town.

Hofstetter, C H (2003) Contextual and mathematics accommodation test effects for English-language learners, *Applied Measurement in Education* 16 (2), 159–188.

Hornberger, N and Chick, K (2001) Co-constructing school safetime: Safetalk practices in Peruvian and south African classrooms, in Heller, M and Martin-Jones, M (Eds) *Voices of authority: Education and linguistic difference,* Westport, CT: Ablex, 31–56.

Hovens, M (2002) Bilingual education in west Africa: Does it work? *International Journal of Bilingual Education and Bilingualism* 5 (5), 249–266.

Johnstone, C J, Altman, J, Thurlow, M and Thompson, S J (2006) *A summary of research on the effects of test accommodations: 2002 through 2004 (technical report no. 45)*, Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kapp, R (2004) 'Reading on the line': An analysis of literacy practices in ESL classes in a south African township school, *Language and Education* 18 (3), 246–263.

Kellaghan, T and Greaney, V (1992) *Using examinations to improve education: A study in fourteen African countries*, Washington, D.C.: World Bank.

Kellaghan, T and Greaney, V (2004) *Assessing student learning in Africa*, Washington, D.C.: World Bank.

Koenig, J A and Bachman, L F (Eds) (2004) *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessments*, Washington, DC: National Academies Press.

Kopriva, R J, Emick, J E, Hipolito-Delgado, C P and Cameron, C A (2007) Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners, *Educational Measurement: Issues and Practice* 26 (3), 11–20.

Larsen, S C, Parker, R M and Trenholme, B (1978) The effects of syntactic complexity upon arithmetic performance, *Learning Disability Quarterly* 1 (4), 80–85.

Lepik, M (1990) Algebraic word problems: Role of linguistic and structural variables, *Educational Studies in Mathematics* 21 (1), 83–90.

Macdonald, C A (1990) Crossing the Threshold into Standard Three, in *Main Report of the Theshold Project*, Pretoria: Human Sciences Research Council.

Mbise, A (1994) Teaching English language reading in Tanzanian secondary schools, in C Rubagumya (Ed) *Teaching and Researching Language in African Classrooms*. Clevedon: Multilingual Matters.

Ministry of Education and Vocational Training (2006) *Education Policy*: Zanzibar.

Mlama, P and Matteru, M (1978) *Haja ya kutumia kiswahili kufundishia katika elimu ya juu (the need to use kiswahili as a medium of instruction in higher education),* Dar es Salaam: Bajita.

Mwinsheikhe, H M (2002, 7–9 January) Overcoming the language barrier: an in-depth study of the Tanzanian secondary school science teachers' initiatives in coping with the English-Kiswahili dilemma in the teaching-learning process, paper presented at the NETREED Conference, University of Oslo.

Mwinsheikhe, H M (2003) Using Kiswahili as a medium of instruction in Tanzanian secondary schools, in Brock-Utne, Desai, Z and Qorro (Eds) *Language of Instruction in Tanzania and South Africa* (LOITASA), Dar es Salaam: E&D.

Myers-Scotton, C (1993) Elite closure as a powerful language strategy: the African case, *International Journal of the Sociology of Language* 103, 143–163.

Ndayipfukamiye, L (1994) Code-switching in Burundi primary classrooms, in Rubagumya, C M (Ed.) *Teaching and researching language in African classrooms,* Clevedon: Multilingual Matters, 79–95.

Ouane, A (2003) *Towards a Multilingual Culture of Education*, UNESCO Institute for Education.

Ogunniyi, M (2005) Cultural perspectives on science and technology education, in Abdi, A and Cleghorn, A (Eds) *Issues in African Education: Sociological Perspectives*, New York: Palgrave Macmillan, 123–140.

Probyn, M (2005) *Learning science through two languages in south Africa*, paper presented at the The 4th International Symposium on Bilingualism, Somerville, MA.

Probyn, M (2006) Language and learning science in south Africa, *Language and Education* 20 (5), 391–414.

Prophet, R and Badede, N (2006) Language and student performance in junior secondary science examinations: The case of second language learners in Botswana, *International Journal of Science and Mathematics Education*, 2009, 7 (2), 235–251.

Rea-Dickins, P, Rubagumya, J and C Clegg (2005) *Evaluation of the Orientation Secondary Class Zanzibar*, a Consultancy Report, University of Bristol: Centre for Research on Language and Education, Graduate School of Education.

Richardson, V (Ed.) (2001) *Handbook of research on teaching* (4th ed), Washington, DC: American Educational Research Association.
Rivera, C and Stansfield, C W (2004) The effect of linguistic simplification of science test items on score comparability, *Educational Assessment* 9 (3 & 4), 79–105.
Roy-Campbell, Z M (1995) Does medium of instruction really matter? The language question in Africa: The Tanzanian experience, *UTAFITI (NS): Journal of Arts and Social Sciences* 2 (1 & 2), 22–39.
Rubagumya, C M (1991) Language promotion for educational purposes: The example of Tanzania, *International Review of Education/Internationale Zeitschrift für Erziehungswissenschaft/Revue internationale l'éducation* 37 (1), 67–85.
Rubagumya, C M (1993) *The language values of Tanzanian secondary school pupils,* unpublished PhD Thesis, University of Lancaster, Lancaster.
Rubagumya, C (Ed.) (1994) *Teaching and Researching Language in African Classrooms*, Clevedon: Multilingual Matters Ltd.
Rubagumya, C M (2003) English medium primary schools in Tanzania: a new 'linguistic market' in education?, in Brock-Utne, Desai, Z and Qorro (Eds) *Language of Instruction in Tanzania and South Africa* (LOITASA), Dar es Salaam: E&D.
Rugemalira, J M (2005) Theoretical and practical challenges in a Tanzanian English medium primary school, *Africa and Asia* 5, 66–84.
Setati, M, Adler, J, Reed, Y and Bapoo, A (2002) Incomplete journeys: Code-switching and other language practices in mathematics, science and English language classrooms in south Africa, *Language and Education* 16 (2), 128–149.
Shorrocks-Taylor, D and Hargreaves, M (1999) Making it clear: A review of language issues in testing with special reference to the national curriculum mathematics tests at key stage 2, *Educational Research* 41 (2), 123–136.
Shorrocks-Taylor, D and Hargreaves, M (2000) Measuring the language demands of mathematics tests: The case of the statutory tests for 11-year-olds in England and Wales, *Assessment in Education: Principles, Policy & Practice* 7 (1), 39–60.
Sireci, S G and Allalouf, A (2003) Appraising item equivalence across multiple languages and cultures, *Language Testing* 20 (2), 148–166.
Sireci, S G and Khaliq, S N (2002) *An analysis of the psychometric properties of dual language test forms*, Amherst: University of Massachusetts.
Sireci, S G, Li, S and Scarpati, S (2003) *The effects of test accommodations on test performance: A review of literature,* Amherst, MA: University of Massachusetts (Amherst).
Solano-Flores, G, Trumbull, E and Nelson-Barber, S (2002) Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities, *International Journal of Testing* 2 (2), 107–129.
Stansfield, C W (2002) Linguistic simplification: A promising test accommodation for LEP students? *Practical Assessment, Research & Evaluation* 8 (7).
Thomas, W P and Coliier, V P (1997) *School effectiveness for language minority students*, Washington DC: National Clearinghouse for Bilingual Education.
Thompson, S J, Blount, A and Thurlow, M (2002) *A summary of research on the effects of test accommodations: 1999 through 2001 (technical report no. 34)*,

Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
Thurlow, M and Bolt, S E (2001) *Empirical support for accommodations most often allowed in state policies (synthesis report 41)*, Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
Uiterwijk, H and Vallen, T (2005) Linguistic sources of item bias for second generation immigrants in Dutch tests, *Language Testing* 22 (2), 211–234.
UNESCO (1953) *The use of vernacular languages in education*, Paris: UNESCO.
UNESCO (2007) *Education for all global monitoring report 2008 – education for all: Will we make it?* Paris: UNESCO.
Wheeler, L J and McNutt, G (1983) The effect of syntax on low-achieving students' abilities to solve mathematical word problems, *Journal of Special Education* 17 (3), 309–315.
Williams, E (1996) Reading in two languages at year five in African primary schools, *Applied Linguistics* 17 (2), 182–209.
Williams, E (2006) *Bridges and barriers: Language in African education and development*, Manchester: St. Jerome Publishing.
Williams, E and Cooke, J (2002) Pathways and labyrinths: Language and education in development, *TESOL Quarterly* 36 (3), 297–322.
Wolf, M K, Kao, J, Herman, J L, Bachman, L F, Bailey, A L, Bachman, P L et al (2008) *Issues in assessing English language learners: English language proficiency measures and accommodation uses*, Los Angeles: University of California, Los Angeles National Center for Research on Evaluation, Standards, and Student Testing.
World Conference on Education for All (1990) *World declaration on education for all and framework for action to meet basic learning needs*, Paris: UNESCO, EFA Forum Secretariat.
Yu, G and Thomas, S (2008) Exploring school effects across southern and eastern African school systems and in Tanzania, *Assessment in Education* 15 (3), 283–305.
Zenisky, A L and Sireci, S G (2007) *A summary of the research on the effects of test accommodations: 2005–2006 (technical report no. 47)*, Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

12 Certifying teachers' foreign language proficiency: developing a performance test for Italian CLIL teachers

Geraldine Ludbrook
Dept of Language Sciences, University of Venice

Abstract

As the provision of Content and Language Integrated Learning (CLIL) is moving increasingly into mainstream education, the call for certified qualification of CLIL teachers is growing. A project is being developed at the University of Venice that seeks to identify the L2 weaknesses and needs of CLIL teachers in Italy, and to respond with specialised pre- or in-service training. The project aims to design a performance test to certify both the L2 competence of CLIL teachers and their knowledge of CLIL methodology. For the purposes of the pilot test, it will focus on the teaching of science through English.

As CLIL is not easily understood as a construct, making the measurement of ability complex, there are many directions for research within this context, which include examining how the interplay of general foreign language proficiency, subject-specific language, the language of classroom interaction, and code-switching contribute to the construction of CLIL science classroom discourse, in addition to what minimum L2 language proficiency is required of the CLIL teacher to effectively handle the methodology needed to implement this approach. This paper will discuss how investigation of the target language use through the qualitative analysis of data from CLIL science classroom observation can help to address some of the main issues that challenge performance test design, such as construct description and test task development.

Introduction

Content and Language Integrated Learning (CLIL) is an educational approach that has evolved in Europe from the new needs for multilingualism set out by the Council of Europe. The term CLIL refers to situations where

subjects, or parts of subjects, are taught through a foreign language with dual-focused aims, namely the learning of content, and the simultaneous learning of a foreign language (Marsh 1994). The approach has been rapidly introduced into mainstream education throughout Europe, yet many issues related to the CLIL teacher remain unaddressed. In particular, the question of CLIL teachers' foreign language proficiency is a little researched area, despite the fact that it is widely considered an essential feature of the success of CLIL: 'One crucial aspect of CLIL should also be spelled out: how good should CLIL teachers' proficiency in the language of instruction be and how could that level be reliably checked?' (Takala 2002:40).

Education authorities throughout Europe have different standards for CLIL teacher foreign language proficiency: the Dutch education authorities recommend at least a B2 level of the CEFR (Common European Framework of Reference), in Hungary a B2–C1 level is required, whilst in Finland the Ministry for Education proposes a C2 level of proficiency, which is also the obligatory level in Poland (Eurydice European Unit 2006:43). Other scholars argue that native speaker skills are a necessary pre-requisite (Smith 2005), while the opinion of one of the leading advocates of the CLIL approach is that 'Teachers do not need to have native or near-native competence in the target language for all forms of delivery, although naturally they need a high level of fluency' (Marsh 2002:11).

Nevertheless, the call for qualified CLIL provision is increasing. In France, additional certification of competence has been required for teachers of a non-language subject teaching in a foreign language since 2003. Germany, too, has introduced additional teacher qualifications for bilingual teaching in some states (Eurydice European Unit 2006:43–44).

As CLIL moves increasingly into mainstream education in Italy, the need for specialised pre-service training and qualification of CLIL teachers is becoming more evident. A project is being developed at the University of Venice to design a performance test to certify the L2 proficiency of CLIL teachers in Italy. The pilot test will be focused on the science classroom, the most common subject taught within the CLIL approach; for the purposes of the initial study, the foreign language used will be English, although aims are to extend the test to other languages to meet the multilingual needs of Italian CLIL.

This paper will examine central issues to be addressed in the development of a language performance test: the definition of the construct to be measured and the specification of the tasks to be administered. An initial analysis of the target language use through a small case study is reported on, and some suggestions for further research are proposed.

Performance language testing

Performance language testing generally tends to follow two main schools of thought. The first is largely a construct-based approach in which performance

is the means by which a language sample is elicited so as to allow evaluation of second language proficiency: McNamara's 'weak' sense of the term (McNamara 1996:43). Test tasks may resemble or simulate real-world tasks, but the real focus of the test is the underlying knowledge and ability that is revealed in the performance, the 'vehicle of assessment' (Messick 1994:14). The meaning of the construct of this kind of performance test 'is tied to the range of tasks and situations it generalises and transfers to' (Messick 1994:15), and provides the criteria used in evaluating task performance.

The construct is generally based on an explicit theory of language and language use, such as the models of communicative language ability developed by Bachman (1990), Bachman and Palmer (1996), and Canale and Swain (1980). Background and topic knowledge, too, are often included in the construct of performance tests for specific purposes, in which test content and test methods stem from an analysis of a specific use situation or context, 'capitalising' on special purpose abilities on the grounds that 'context-based tests may provide more useful information than general-purpose tests when the goal is to make situation-specific judgments about subjects' communicative language ability' (Douglas 1997:18).

The second theory is the task-based approach to performance testing – McNamara's 'strong' sense of the term (McNamara 1996:43) – in which the fulfilment of the test-task is the 'target of assessment', and the second language is the 'medium' of the performance (Messick 1994:14). The test tasks simulate or replicate real-world tasks and the criteria used for evaluation of task fulfilment are based on real-world criteria. In its most pragmatic form, this approach may make no recourse to theoretical models of language use in the definition of the test construct, relying instead on a close analysis of the target language use: 'Task-based assessment does not simply utilize the real-world task as a means for eliciting particular components of the language system which are then measured or evaluated; on the contrary, the construct of interest in task-based assessment is performance on the task itself' (Brown et al 2002, cited in Bachman 2002:455). To what extent a test of this kind can actually provide a basis for interpretations beyond the task or task context, the question of justifying inferences from test performance (McNamara 1996:17), is precisely one of the challenges test designers working in this approach must address.

Bachman (2002) takes up this challenge and, whilst fully aware of the limitations of generalisability and extrapolation offered by atheoretical task-based testing, proposes that test design take into consideration both construct definition and task specification, attempting to define task characteristics as closely as possible to the facets of the assessment in question on the basis of both the analysis of the target language use domain and either an existing framework or a framework developed *ad hoc* for the test. Bachman also refers to the construct definition of the specific areas of language abilities

to be assessed, suggesting that the construct may include several separate components or involve a global definition. The construct can be rooted in a theoretical model, or based on a course syllabus, or defined from a needs analysis of the target language use domain; it may attempt to measure all or parts of any of these aspects of the construct. 'Planned integration of both tasks and constructs in the way they are designed, developed and used' may provide test developers with 'the full range of validation arguments that can be developed in support of a given inference or use' (Bachman 2002:471).

Defining an appropriate *a priori* construct for a performance language test for Italian CLIL teachers, and considering the test tasks to be designed, will therefore require a careful analysis of the target language use domain. The next section of this paper will examine the target language use in the Italian CLIL science classroom, drawing both on the literature and on a small case study recently carried out in the Italian context.

Background to the study

CLIL methodology

The CLIL approach is rooted in a dual focus on language and content, a vehicular use of the foreign language. The approach draws heavily on strategies taken from models of content-based teaching (Brinton, Snow and Wesche 1989). The 'sheltered approach' to subject matter teaching used in content-based instruction involves a wide range of scaffolding strategies to communicate meaningful input in the content area, as well as adapting the language of texts or tasks and using methods such as visuals, graphic organisers, or co-operative work to make instruction more accessible to students with low levels of L2 proficiency.

Another resource that is drawn on in CLIL methodology is task-based teaching. In this method, teachers 'interactionally support task performance in such a way as to trigger processes such as the negotiation of meaning and content, the comprehension of rich input, the production of output and focus on form, which are believed to be central to (second) language learning' (Van Avermaet, Colpin, Van Gorp, Bogaert and Van den Branden 2006:175). In task-based learning classrooms, the teacher tends to ignore language errors and focus more on the real aim of the task. In this way, the teacher 'puts the initiative for solving comprehension problems, running the conversation and initiating the topic into the hands of the learner' (Van Avermaet et al 2006:175).

CLIL teachers therefore have to possess a level of L2 competence that will enable them to implement CLIL methodology. But the approach calls for further considerations. CLIL teachers are also required to devise and modify materials and tasks that will permit the learners' access to the content subject in the foreign language. Moreover, CLIL teachers and learners are

all simulating an L1 classroom situation in a language that is foreign to both groups and, although a strong L2-only policy is generally advocated, the question must be raised as to what role the teacher's and learners' own L1 will play in the CLIL classroom and what use CLIL teachers can make of it to enhance communication. In addition to the other language skills mentioned above, CLIL teachers need sufficient specialist language knowledge, of both genre and lexis, to teach the content subject in the foreign language.

CLIL in the Italian context

In 1999, education reform in Italy paved the way for a more widespread introduction of the CLIL approach in mainstream education. State schools were given greater autonomy to introduce and develop different forms of teaching that more closely met the needs of their students. Amongst these was the possibility to teach content subjects in a foreign language. Another innovation was the introduction of more flexible forms of teaching, in particular the concept of teaching modules, which may be of variable length, from a few hours to several months, and may have a cross-curricular nature (see Coonan 2002:43–44). The focus on flexible language instruction was further reinforced by *Progetto Lingue 2000*, a project of the Italian Ministry for Education to improve the quality of foreign language teaching in the state school system (MPI 2000). CLIL is currently delivered in over 100 pilot projects throughout the country, generally in a modular format. Although English is the most popular foreign language, all the Italian projects have a strong multilingual policy, and CLIL in French, German and Spanish is strongly encouraged (Eurydice European Unit 2006:34).

Since the early 1990s, Italian education authorities have organised projects for CLIL teacher development. In the Veneto region, for example, the University of Venice has run training courses in CLIL methodology for in-service teachers in collaboration with regional education authorities since 2002, and is working to introduce the training of pre-service teachers (Coonan 2004a). In addition, pan-European CLIL projects, under the Socrates scheme, have funded teacher mobility programmes for language and subject teachers alike, to improve their language skills or to follow CLIL teacher training courses abroad (Coonan 2002:107–108).

There are some content teachers who teach CLIL on their own; however, in Italy CLIL is mainly provided through a teaching team of subject and foreign language teachers. In the Italian CLIL classroom, the teaching partnership seems to be characterised by features of the complementary/supportive teaching team, defined by Maroney (1995) as one in which 'one teacher is responsible for teaching the content to the students, while the other teacher takes charge of providing follow-up activities on related topics or on study skills'. In some cases, the collaboration takes place before the lesson and the

content teacher manages the lesson on his/her own. More commonly, in addition to shared preparation, both teachers are always present in the classroom at the same time (see the examples reported in Coonan 2004b).

The case study

The principal objective of the case study was to pilot a classroom observation checklist developed for the purpose of identifying the language features involved in CLIL teaching performance as a tool for defining a framework for the construct of the test. A second goal of the study was to begin to examine what minimum level of language proficiency a content teacher working in a team-taught CLIL context might need to have.

The school chosen for the case study was an Istituto Tecnico Statale, a technical secondary school that trains students for employment in the sectors of trade, tourism and surveying. It was chosen as the context for the case study because English-language CLIL in the science classroom has been implemented here for several years, generally in the first two years of secondary school with students aged 14–16.

The class observed was made up of 20 students aged 15: four boys and 16 girls. The students had already received science instruction in CLIL the previous year with the same teachers and were therefore familiar with the procedures and classroom rules regarding the use of English, as well as with pair and group work activities.

Methodology

A qualitative approach was adopted in the study, incorporating methods of data collection to build up as rich a picture as possible of the CLIL learning and teaching environment. Semi-structured interviews were carried out separately with content and EFL teachers both to obtain background information on the classes to be observed and to put together a portfolio of the subject teacher's English competences.

Four CLIL lessons were observed and audio recordings made, which were then transcribed and coded for a close study of the CLIL classroom discourse and, in particular, the CLIL teacher's use of language. As the observations of the CLIL classrooms were exploratory and diagnostic, an observation checklist was chosen as a useful tool as a framework for the observation. Often used to provide a sampling frame to classroom observation (see, for example, Montgomery 2002) this instrument has also been used for both *a priori* and *a posteriori* analysis of output in speaking test tasks (O'Sullivan, Weir and Saville 2002).

Two checklists devised for classroom observation of non-native English speaking teachers were examined in the preliminary stages of the development

of a similar tool for Italian CLIL classrooms. De Graaff, Koopman, Anikina and Westhoff (2007) report on the development of an observation tool based on principles from second language pedagogy. The checklist covers several aspects of CLIL methodology: focus on form, focus on meaning and different kinds of scaffolding. It does not, however, look at the fields of general language proficiency, subject-specific language or classroom management, as the specific aim of De Graaff et al's study is to detect effective CLIL pedagogy.

Closer to the aims of the Italian testing project, albeit in a non-European context, Elder (1993) illustrates an observation schedule developed 'to assess the English language proficiency of non-native speaker graduates training as secondary mathematics and science teachers' in Australian schools (Elder 1993:235). The schedule contains features of both language and language-related behaviour based on the literature of classroom communication, considered crucial for effective teacher performance and revised to include only those features which were found 'to discriminate among non-native speaker teachers' (Elder 1993:237). The schedule was produced for use by teachers of mathematics and science, so was formulated to be meaningful to non-language experts and designed to be used during a 15-minute observation of teacher performance.

As Elder's 1993 schedule contained some of the main categories of language features considered relevant to the Italian CLIL context, it was decided to use this schedule as a starting point for the CLIL classroom observations. A group of Italian experts in CLIL methodology, teacher trainers, CLIL teacher trainers, and trainee teacher supervisors was asked to indicate what aspects of the original schedule they considered to be important features of the Italian CLIL classroom. Their evaluations were then incorporated into a revised version of the schedule that attempted to focus more precisely on the foreign language needs of the Italian CLIL teacher. An additional section was added to the schedule, which took into consideration code-switching, intended here as any kind of alternation between L1 and L2, not specifically switching, borrowing or mixing. Although L2-only interaction is encouraged, the effective use of L1 is an important feature in CLIL classroom discourse (see, for example, the studies by Butzkamm 1998 and Nikula 2005 for further research on code-switching practices in CLIL classrooms). The two descriptors added concerned the teacher's effective use of L1/L2 code-switching and the teacher's encouragement of effective code-switching by the students. The CLIL observation schedule used in this pilot study can be found in the Appendix.

In addition to teacher interviews and classroom observation, various documents used in the CLIL module were examined: handouts prepared by the content teacher and used by the EFL teacher to prepare students for the CLIL module, and the tasks set for students during the module. The

end-of-module test was also looked at. The test combined multiple-choice and true/false items with open-ended questions and was marked by both science and EFL teachers, with separate grades given for content and language.

A group interview with six teachers implementing the CLIL approach in the school in question was organised to discuss the specific questions of CLIL teachers' language needs and levels, drawing on the experience of both foreign language and content subject teachers. Coonan (2007) has conducted considerable research on the 'insider' view of the CLIL classroom, working with subject and language teacher teams implementing the approach in Italian classrooms, to record their perceptions of the CLIL classroom. Her results show that, due to a high degree of teacher awareness, useful information can be gleaned from CLIL teachers' experience in the classroom. In this study, the group was made up of three content teachers (two science teachers and one history teacher), two Italian EFL teachers and one native speaker 'conversation' teacher. All six had at least two years' experience of the CLIL approach and all had completed, or were currently following, CLIL methodology training at the University of Venice.

The CLIL science teacher

A semi-structured interview was carried out with the science teacher before the observations with the purpose of gathering data on his CLIL experience and to put together a portfolio of his English language background. Male, an Italian national and native speaker, PP has a degree in chemistry and has been teaching science at secondary school for over 26 years. After studying English at school for eight years, he then conducted most of his university studies using English language textbooks. Since then, his use of English has largely been limited to personal use (mainly television and film), consultation of online scientific journals, and attendance at European conferences. He has, however, been involved in Comenius exchange programmes, working with visiting teachers from schools in Wales and Lithuania, even though he has not spent time in either country.

PP is a strong advocate of the CLIL approach and has been instrumental in introducing it into the school. He completed a CLIL training course at the University of Venice and was involved in a research project involving CLIL teachers in Italy. He uses internet resources to provide material for his CLIL module, including Massachusetts Institute of Technology videos of science lessons.

In an attempt to establish PP's level of English proficiency two different tools were used. Firstly, he was asked to provide an evaluation of his level of language competence using the Common European Framework of Reference self-assessment grids (see CEFR, Tables 2 and 3, Council of Europe 2001:26–29). He placed himself within the B1 level for all skills, with the exception

of reading comprehension, which he evaluated as B2. Secondly, PP assessed his English proficiency using the DIALANG diagnostic language tests. In reading, grammar and vocabulary, his results were at the C1 level, whereas his listening comprehension score was slightly lower at the B2 level.

When asked about his perception of his strengths and weaknesses, PP spoke of the amount of care with which he prepares his teacher-fronted laboratory lessons, with which he feels confident as, familiar with his subject, he can concentrate on his use of English. His greatest difficulties arise in unplanned interaction, the unpredictable lexis that he might require during the lesson to respond to student requests for information.

PP also outlined the structure of the CLIL module planned. It was to last 18–20 hours and would be delivered in the last five weeks of the school year. The students would first be made familiar with some of the vocabulary to be used by the EFL teacher in the English lessons. Then a series of four lessons would be held in the physics laboratory. In these teacher-fronted lessons, PP would carry out demonstrations and experiments related to the theme of the module. The next four lessons would be group work held in the multimedia laboratory. The students would work in pairs on a task that involved retrieving information from the internet. The students would then prepare a PowerPoint presentation of the completed task.

PP also provided insight into the role of the two teachers in the CLIL teaching team. He explained that he chose the materials to be used in class, mainly from the internet for its greater flexibility, adapting them slightly, mainly by reducing the length. He then passed the material on to the EFL teacher who devised exercises to be used in her EFL lessons. He stressed that the role of the EFL teacher in the CLIL classroom is that of providing language support, intervening when she sees students in difficulty, or when a lexical problem occurs.

The EFL teacher

The formal interview with the EFL teacher took place after the observations, although several informal conversations had taken place before and during the observations. She had team-taught CLIL with PP and another science teacher at the school for two years. The teacher confirmed that during the CLIL module, all her EFL lessons were given over to preparation of the CLIL science lessons. Her assessment of the class's English language skills was that they had an overall good level of comprehension with varying levels of written and oral production.

With regard to PP's language needs, the EFL teacher saw the shift from his working within his subject (what she called ESP) to other registers, such as class management, as being his greatest difficulty, as he lacked the 'lexical richness and flexibility' to answer student questions with ease. She also

mentioned the issues of intonation and pronunciation as being areas of difficulty for this particular teacher.

Observing the CLIL classroom

The first two lessons observed were held in the physics laboratory and involved PP explaining a process and illustrating it through a series of practical demonstrations. The topic of the module was electrostatics. These lessons were science teacher-fronted activities, while the EFL teacher stood at the whiteboard providing written support (for example, irregular verbs, specialist lexis) and occasionally intervening orally.

A further two lessons observed were held in the multimedia laboratory. In these lessons, the students worked in pairs retrieving information from the internet to respond to a series of questions they had been assigned while the teachers monitored and assisted them.

Classroom management

A first observation was that in both lesson types (teacher-fronted and group work), most of the classroom management, such as introducing the lesson, setting up activities, and disciplining the students, was carried out by the EFL teacher. The EFL teacher opened lessons with revision of content material dealt with in the previous lesson; she also closed lessons with instructions for the next meeting. Many of the descriptors in the CLIL classroom observation schedule could therefore not be related very closely to the CLIL content teacher's performance.

The fact that much of the classroom management was carried out by the EFL teacher meant that it was hard to evaluate the validity of the CLIL schedule on the basis of these observations. Further observations will be needed to verify whether this division of tasks by the EFL and content teachers is common to many teaching teams or whether it was specific to this particular pair.

Using subject-specific language

During the experiments in the teacher-fronted lessons, a useful sample of subject-specific language was recorded. Regarding the knowledge of subject-specific terms, PP appeared to have good control of the specialist lexis of electrostatics. Moreover, not only did he pronounce specialist terms clearly and correctly, he also corrected the EFL teacher's mispronunciation of specialist terms. In addition, he also helped the EFL teacher in the correct spelling of subject-specific words.

PP linked ideas using very simple connectors, such as, for example, *if* and *so*. The students appeared to have little difficulty in understanding the simple demonstrations of the principles of electrostatics. In addition PP's

experiments were scaffolded by the handouts distributed to students and to which the EFL teacher invites students to refer during the lesson.

The section of the observation schedule regarding the use of subject-specific language seems to correspond to the content teacher performance observed.

Using L1 and L2

In their respective interviews, both content and EFL teacher claimed that they attempted to maintain a strong L2 policy in the CLIL classroom, using Italian only as a last resort after several attempts to reformulate have been made. Both teachers repeatedly issue instructions to students to use English.

In the classroom interaction recorded, Italian was used in four different ways:

1. by students in response to an explicit request for the Italian translation by the content teacher:

   Content Teacher (CT): This is the box full of Styrofoam chips. OK? It what is Styrofoam in Italian?

   Student (S): *Polistirolo.*

2. in 'private' interaction between the content and EFL teachers:

   CT: This is a bakelite rod. OK. [*sotto voce* to EFL Teacher who is writing the term on the board] *bakelite con kappa*. [bakelite written with a k]

3. by the content teacher in response to a student question in Italian:

   S: And what mean 'drive' in this case?

   CT: Uh you'll see you'll see.

   S: *Qual'è il significato di 'drive'?* [What does 'drive' mean?]

   CT: *Immergere.*

4. the content teacher provides translation of previous statement in English.

   CT: Bakelite rod. Bakelite is *bachelite* in Italian.

The descriptors seem to capture the use of L1/L2 code-switching in the lessons observed. If further observations confirm that the different ways in which Italian is used by the content teacher are systematic to CLIL science classrooms, the descriptors might be articulated to take this into consideration. In the case of this study, the interactions recorded do not seem to correspond to the explicit policy advocated by the teachers in their interviews.

General language proficiency

I have left to last the most complex section of the CLIL teacher's performance, that of general language proficiency. This is the section in which issues

such as intelligibility, fluency and accuracy come into play, requiring clear definitions if qualitative judgments are to be made.

Native-speaker language proficiency does not seem to be the model in the CLIL approach. Many countries have indicated minimum levels of teacher foreign language proficiency using the Common European Framework of Reference, and educationalists promoting the CLIL approach have explicitly advocated the non-native speaker model (see Graddol 2005 and Marsh 2002).

Performance on language tests is typically judged with reference to a native speaker ideal. Some scholars have, however, challenged the concept that the native speaker is an appropriate model of English for language testing, and teaching, outside Kachru's (1990) 'Inner Circle' (see work by Elder and Davies 2006, House 2002, Jenkins 2006, Seidlhofer, Breiteneder and Pitzl 2006, Taylor 2006). CLIL would seem to be a clear example of English used as a lingua franca in the classroom, albeit between non-native speakers sharing the same first language. A discussion of the issues involved in terms of what model is to be used in the evaluation CLIL teacher performance will be necessary.

Intelligibility of expression

The concept of intelligibility is a complex one involving both speaker and hearer. The hearer's understanding may depend on whether the speaker's accent is familiar, on the hearer's inferencing skills and on knowledge of the topic. The speaker's production may depend on pronunciation (stress, rhythm, intonation, voice quality and sounds), delivery (hesitations, uncertainty, volume), grammar, sound symbol relationships.

Although a discussion of how the processes by which understanding is achieved in ELF interaction may be qualitatively different from native speaker-based interaction will be necessary (see Brown and Lumley 1998, Elder and Davies 2006, Han and Singh 2007, Pickering 2006), for the purposes of this initial study, intelligibility of the content teacher in the CLIL classroom in this study is considered in terms of the researcher's perception of the Italian student listeners' understanding based on classroom observations. In other words, whether the teacher was intelligible to students with limited English proficiency working with a teacher whose accent, pronunciation and delivery are familiar, as is the topic of the lesson.

In the observations recorded, the students seem to have little difficulty in understanding the content teacher working in English. Their problems appear to be in understanding the subject content, or specialist lexis, rather than the science teacher's delivery.

PP stresses important words and makes easy transition from one experiment to the next. His use of non-verbal strategies would seem to be appropriate to the science classroom situation.

Fluency and flexibility of expression

In the literature, fluency as a descriptor for oral language performance has been defined in various ways. Crystal, for example, defines fluency as 'smooth, rapid, effortless use of language' (Crystal 1987:421); Brumfit talks of 'natural language use whether or not it results in native speaker-like language comprehension or production' (Brumfit 1984:56). Chambers (1997) looks at recent research into definitions of fluency, including speech rate and pauses, and compares the difference in fluency in native and non-native speakers.

In the presentation phase of the lessons, PP maintains a fluent description of the experiments and procedures he is carrying out. Less fluency is noted in the teacher–student interactions during the monitoring of students in group work activities.

Turning to the descriptor regarding flexibility of expression, PP appears to have little lexical flexibility and variety, and often tends simply to repeat his previous statement with little variation on the original, as in this example:

CT: Like charges repel each other. For example positive and positive they are like charges or negative and negative. Like charges are of the same sign. OK? Both positive or both negative.

Accuracy of expression

In the testing of oral performance, accuracy is generally perceived as being based on grammatical correctness, and is often contrasted with fluency. Within the CLIL approach, however, as in other forms of content-based instruction in which there is a dual focus on language and content, fluency is favoured over accuracy and native-speaker competence is not aimed for (see, for example Marsh 2002:36). Student errors are generally only corrected when communication breakdown or misunderstanding occurs. This explicit policy regarding accuracy will therefore also affect the model the CLIL content teacher is expected to provide, especially in the co-taught classroom.

Assuming, for the purposes of this initial study, grammatical correctness as a definition of accuracy, PP makes very few errors in the teacher-fronted lessons. Some examples of grammatical and syntactical inaccuracy are:

CT: I'm going to do some experiments and then I'm going to commenting them with you.

CT: The electroscope told that the Cavendish hemisphere is charged.

CT: When I shake fast the rod . . .

CT: What kind of forces are they between the paper and the white board?

None of these inaccuracies caused any misunderstanding with the students.

Planning, monitoring and repair

PP has evidently planned his teacher-fronted lessons with great care. In the interview, he claims each lesson takes 3 to 4 hours' preparation. When he

does make mistakes or slips, he is very often able to correct himself, whether they are minor grammatical slips or errors of pronunciation.

Conclusions and further research

This initial analysis of the observation schedule seems to indicate that it is a useful tool in the observation of the CLIL science teacher's performance. Most of the descriptors seem relevant and capture salient moments of the classroom interaction.

In future research, validation of the checklist will continue with repeated observations of science classrooms and through focus group discussions with other teachers using the CLIL approach in Italy aimed at further clarifying and refining the checklist. This framework drawn from the target language use domain will form the basis for the construct underlying the test, and guide the construct-based scoring criteria used for performance evaluation.

The case study also provided insight into the tasks that might be designed for the performance test, operationalising the construct. The teacher needs to be able to prepare and deliver teacher-focused presentation of subject-specific material, with the aid of practical demonstrations, board work, and written handouts; to set up and monitor pair and group work task-based activities, interacting with the students on issues regarding both content and language; to evaluate student performance, both oral presentation of group work tasks and written test production. Establishing the nature of the test tasks and defining the task characteristics will require careful consideration of what degree of authenticity and interactivity is desired and can be achieved in a performance test simulating a classroom situation.

As regards the issue of starting to establish what minimum level of language skills a teacher implementing CLIL in Italy needs to possess, from this brief case study it would appear that in the science classroom, a content teacher with a language profile that ranges between a B2–C1 level of general English competence seems to be able to function in the presentation phase of the CLIL lesson. His delivery of prepared material shows considerable accuracy – grammatical, syntactical, lexical and of pronunciation – and he is able to monitor and correct both his own errors, and those of students. When dealing with subject-specific terms, he is also able to monitor the students' and the EFL teacher's speech. Fluency and pronunciation also appear to be appropriate for the level of the students being taught.

However, it would also seem that this level of general English competence is insufficient for some aspects of the teaching performance. The content teacher depends heavily on the presence of the EFL teacher for the phases of the lesson that require more flexible and interactive language use, such as opening and closing the lesson, and setting up activities. Yet his familiarity

with his subject and teaching experience provide him with tools that assist his teaching performance in English.

Some of the questions that future research will therefore have to explore are what factors contribute to task fulfilment: language proficiency or teaching strategies/classroom competence, and how teaching experience can be separated from language performance in the test situation in order to avoid construct-irrelevant variance, a major threat to construct validity in performance tests.

Acknowledgements

The author would like to thank the CLIL teachers at Istituto Tecnico Statale per il Commercio, il Turismo e per Geometri Girardi, Cittadella, for their kind collaboration.

References

Bachman, L F (1990) *Fundamental considerations in language testing*, Oxford: Oxford University Press.

Bachman, L F (2002) Some reflections on task-based language performance assessment, *Language Testing* 19 (4), 453–476.

Bachman, L F and Palmer, A S (1996) *Language testing in practice*, Oxford: Oxford University Press.

Brinton, D M, Snow, M A and Wesche, M B (1989) *Content-based second language instruction*, Mass: Heinle & Heinle.

Brown, A and Lumley, T (1998) Linguistic and cultural norms in language testing: a case study, *Melbourne Papers in Language Testing* 7 (1), 80–96.

Brumfit, C J (1984) *Communicative Methodology in Language Teaching: The Roles of Fluency and Accuracy,* Cambridge: Cambridge University Press.

Butzkamm, W (1998) Code-switching in a bilingual history lesson: The mother tongue as a conversational lubricant, *International Journal of Bilingual Education and Bilingualism* 1 (2), 81–99.

Canale, M and Swain, M (1980) Theoretical bases of communicative approaches to second language teaching and testing, *Applied Linguistics*, 1 (1), 1–47.

Chambers, F (1997) What do we mean by fluency?, *System* 25, 535–544.

Coonan, C M (2002) *La lingua straniera veicolare*, Torino, UTET.

Coonan, C M (2004a) La formazione avanzata a livello universitario dei docenti CLIL in Italia. Retrieved 17.5.2006. <gold.indire.it/nazionale/content/index.php?action=read_cnt&id_cnt=5967>.

Coonan, C M (2004b) *CLIL: un nuovo ambiente di apprendimento*, Venice: Libreria Editrice Cafoscarina.

Coonan, C M (2007) Insider Views of the CLIL Class Through Teacher Self-Observation and Introspection, *The International Journal of Bilingual Education and Bilingualism*, 10 (5), 625–646.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*, Cambridge: Cambridge University Press.

Crystal, D (1987) *The Cambridge Encyclopaedia of Language*, Cambridge: Cambridge University Press.

De Graaff, R, Koopman, G J, Anikina, Y and Westhoff, G (2007) An observation tool for effective L2 pedagogy in content and language integrated learning (CLIL), *The International Journal of Bilingual Education and Bilingualism*, 10 (5), 603–624.
Douglas, D (1997) *Testing Speaking Ability in Academic Contexts – Theoretical Considerations*, TOEFL Monograph Series 9, Princeton, NJ: ETS.
Elder, C (1993) How do subject specialists construe classroom language proficiency?, *Language Testing* 10 (3), 235–253.
Elder, C (1994) Performance testing as a benchmark for LOTE teacher education, *Melbourne Papers in Language Testing* 3 (1), May 1994, 1–25.
Elder, C (2001) Assessing the language proficiency of teachers: are there any border controls?, *Language Testing*, 18 (2), 149–170.
Elder, C and Davies, A (2006) Assessing English as a Lingua Franca, *Annual Review of Applied Linguistics* 26, 282–301.
Eurydice European Unit (2006) *Content and Language Integrated Learning (CLIL) at School in Europe*. Retrieved 28.05.2006. <www.eurydice.org/portal/page/portal/Eurydice/showPresentation?pubid=070EN>
Graddol, D (2005) CLIL debate questions and answers, *Guardian Weekly*, Wednesday April 20 2005. Retrieved 18.08.2007 <www.guardian.co.uk/theguardian/2005/apr/20/guardianweekly.guardianweekly13>
Han, J and Singh, M (2007) World English Speaking (WES) student-teachers' experiences of schools: curriculum issues, transnational mobility and the Bologna Process, *Transnational Curriculum Inquiry*, 4 (1). Retrieved 13.05.2008. <http://nitinat.library.ubc.ca/ojs/index.php/tci>
House, J (2002) Developing pragmatic competence in English as a lingua franca, in Knapp, K and Meierkord, C (Eds), *Lingua franca communication*, Frankfurt a. M. et al.: Lang.
Jenkins, J (2006) The spread of EIL: A testing time for testers, *ELT Journal*, 60 (1), 42–50.
Kachru, B (1990) World Englishes and applied linguistics, *World Englishes* 9 (1), 3–20.
Maroney, S (1995) *Team Teaching*. Retrieved 12.02.2008. <www.wiu.edu/users/mfsam1/TeamTchg.html>
Marsh, D (1994) *Bilingual Education & Content and Language Integrated Learning*, Paris: International Association for Cross-cultural Communication, Language Teaching in the Member States of the European Union (Lingua) University of Sorbonne.
Marsh, D (Ed.) (2002) *CLIL/EMILE. The European Dimension*, Finland: UniCOM Continuing Education Centre, University of Jyväskylä.
McNamara, T (1996) *Measuring Second Language Performance*, Harlow: Longman.
Messick, S (1994) The interplay of evidence and consequences in the validation of performance assessments, *Educational Researcher*, 23 (2), 13–23.
Ministero della Pubblica Istruzione (2000) *Progetto Lingue 2000*, Roma: Gruppo Lingue Straniere, MPI.
Montgomery, D (2002) *Helping Teachers Develop through Classroom Observation*, 2nd ed, London: David Fulton Publishers.
Nikula, T (2005) English as an object and tool of study in classrooms: Interactional effects and pragmatic implication, *Linguistics and Education* 16 (1), 27–58.
O'Sullivan, B, Weir, C J and Saville, N (2002) Using observation checklists to validate speaking-test tasks, *Language Testing* 19 (1), 33–56.

Pickering, L (2006) Current research on intelligibility in English as a lingua franca, *Annual Review of Applied Linguistics*, 26, 219–233.
Seidlhofer, B, Breiteneder, A and Pitzl, M-L (2006) English as a Lingua Franca in Europe: Challenges for applied linguistics, *Annual Review of Applied Linguistics* 26, 3–34.
Smith, K (2005) Is this the end of the language class?, *Guardian Weekly*, January 21, 2005. Retrieved 11.11.2006. <education.guardian.co.uk/tefl/teaching/story/0,15085,1394830,00.html>
Takala, S (2002) Positioning CLIL in the Wider Context, in Marsh, D (Ed.) *CLIL/EMILE. The European Dimension*, Finland: UniCOM Continuing Education Centre, University of Jyväskylä, 40–42.
Taylor, L (2006) The changing landscape of English: implications for language assessment, *ELT Journal*, 60 (1), 51–60.
Van Avermaet, P, Colpin, M, Van Gorp, K, Bogaert, N and Van den Branden, K (2006) The role of the teacher in task-based language, in Van den Branden, K (Ed.) *Task-Based Language Education*, Cambridge: Cambridge University Press, 175–196.

APPENDIX

CLIL classroom observation schedule

General language proficiency

1. Intelligibility of expression

1.1 pronounces words/sounds clearly
1.2 utters sentences clearly, with suitable rhythm and intonation
1.3 stresses important words/ideas
1.4 clearly marks transition from one idea/lesson stage to the next, using words such as *so*, *now*, *right*
1.5 uses appropriate facial expressions, gestures, body movement

Fluency and flexibility of expression

1.6 speaks at a speed appropriate to the level of the class
1.7 speaks fluently, without too much uncertainty
1.8 can express ideas in different ways: rephrasing, elaborating, summarizing, exemplifying

Accuracy of expression

1.9 grammar of spoken and written language is generally accurate
1.10 uses correct spelling and punctuation in board work

Planning, monitoring and repair

1.11 plans what is to be said and the means to say it, exploiting any resources available
1.12 uses circumlocution and paraphrase to cover gaps in vocabulary and structure
1.13 backtracks when a difficulty is encountered and reformulates
1.14 corrects own slips and errors if s/he becomes aware of them or if they have led to misunderstandings

2. Using subject-specific language

2.1 demonstrates knowledge of subject-specific terms
2.2 pronounces specialist terms clearly
2.3 uses specialist terms judiciously, writing on board when necessary
2.4 makes clear the connection between ideas, stressing link words *if*, *since*, *in order*
2.5 explains concepts and processes in ways appropriate to the level of the class, using simple language and familiar/concrete examples
2.6 explains diagrams, models, graphs clearly
2.7 links new information to the students' previous knowledge

3. Using the language of classroom interaction

3.1 poses questions to check understanding of previously learned material/ new information
3.2 grades questions appropriately for the level of the class and the learning task: simpler to more complex; closed/open
3.3 responds appropriately to students' questions, requests for assistance
3.4 deals effectively with wrong answers, non-response, using scaffolding techniques such as requests for clarification and recasts
3.5 gives clear instructions for activities
3.6 makes effective use of teaching materials

4. Using L1 and L2

4.1 makes effective use of L1/L2 code-switching, clarifying rules with students
4.2 encourages students' effective use of L1/L2 code-switching

13 Common reference for the teaching and assessment of 'Intercultural Communicative Competence' (ICC)

Denise Lussier
McGill University, Montreal

Abstract

This article is based on the development of a conceptual framework of 'intercultural communicative competence' (ICC) (Lussier 1997, in press) which was field tested and validated among 2,000 English and French speaking young adults, minority and majority ethnic groups from two major bilingual cities in the provinces of Quebec and Ontario, Canada.

The focus of the study is on the development of positive cultural representations as an essential component of ICC for better understanding of other cultures. It is related to positive attitudes expressed via behaviours and practices conveying openness to others and other cultures and to the negative attitudes which are expressed via behaviours rejecting others. It builds on current research considering education as the entry to culture (Bruner 1996) and, more specifically, on the dimensions of language teaching and learning, as a discipline which embodies by nature the presence of another culture and contacts with others.

The article is threefold. It presents: 1) a conceptual framework of ICC and the description of the three domains of reference – knowledge, skills and attitudes defined in terms of competence; 2) guidelines for the assessment of ICC; 3) examples from a survey on cultural representations of ICC among European teachers, and from guidelines on the assessment of ICC in terms of levels of proficiency and scale of descriptors from the domain of 'existential competence' based on the affective and psychological domains.

Major issues in language curriculum, teaching and assessment

Globalisation, with the mobility of people, worldwide communication and new technologies, is a new reality. It brings socio-economical and political changes. All nations find it difficult to deal with such phenomena. Even cultural and religious borders overlap geographical borders. It brings the question of self-identity in plurilingual and pluricultural societies. These new issues create misunderstanding and new confrontations. In such a context, factual information, formal explanations and cognitive reasoning are no longer sufficient; perceptions of others, representations of other cultures and religions, as elements of the affective domain, become a new dimension to explore which is more difficult to define. Such findings give prominence to the interrelation of 'language', 'thought', 'culture' and 'identity' in the development of each individual. It is a key issue which needs to be addressed, hence the domain of 'intercultural communicative competence'.

For the purpose of this article, the focus is only on the construct of cultural representations in the context of language teaching and evaluation. Many questions come to mind. How is language connected to culture and thought? How are cultural representations, positive and negative, constructed? How can we explain common or different cultural representations within the same country and with other countries? How can we define intercultural communicative competence? How can we conduct the teaching/learning and assessment of such competencies? To answer these questions, empirical studies are needed to help educators and researchers to understand the phenomenon and find a common reference to generate a model of intercultural communicative competence (ICC).

This article is the result of a major Canadian empirical study conducted in the English and French speaking communities of Montreal and Ottawa to explore the interrelations between these key issues (Lussier, Auger, Clément and Lebrun-Brossard 2002–2008). The general aim was to analyse the '*development of cultural representations, ethnic identity and intercultural competence'* among 1,500 young adults in the Canadian linguistic context. The first specific objective was to detect positive cultural representations (linked to xenophilia and openness to other cultures) and negative cultural representations (linked to xenophobia, rejection or intolerance of other cultures). It seeks to investigate how cultural values are mediated and to document the role of a cultural mediator in reducing misunderstandings and managing conflicts. The second specific objective was to validate the conceptual framework developed to study the complex phenomenon of intercultural communicative competence (Lussier 1997, in press) taking into account influential factors linked to cultural representations and ethnic identity.

Statement of the problems

It is increasingly accepted to view language teaching and learning as a means of discovering another culture. It does embody, by its nature, the presence of another culture and the contact with alterity. It generates new 'cultural representations' and requires an important part of mediation when interacting with members of other cultures. In this perspective, the development of intercultural competence is now considered as important as the development of the linguistic, sociolinguistic and discursive components of communicative communication (Kramsch 1998, Byram 1997, Lussier 1997, Galisson 1991).

Cultural representations are defined as mental and public representations inhabiting a given social group. Beliefs, intentions, preferences and values are mental representations specific to individuals and societies. Signals, utterances, texts, discourse and pictures are public representations as they are expressed orally, visually or in writing (Sperber 1996:24, Bourdieu 1982). These representations influence our thoughts. They shape our vision of the world. They have an impact on the construct of self-identity.

In this perspective, when teaching languages, we make the assumptions that: 1) educators underestimate the intercultural dimension in the learning process; 2) educators, too often, limit teaching to public representations such as stereotypes, artefacts, etc.; 3) the development of intercultural communicative competence is not considered to be a crucial and essential component of communicative competence. We also argue that the concepts of *language – thought – culture* are interrelated and that 'we cannot teach a language for long without coming face to face with social contexts and cultural factors which have bearing on language' (Stern 1983:191).

Assumptions about language: what is 'to communicate'?

One of the complaints concerning language definitions is that they let us believe that there is a neutral link between the speaker and language. There is also a whole set of attitudes and emotions which renders language analysis superficial if studied as a simple tool with which to communicate. It is essential to explain and account for social communication in its full complexity (Calvet 1999). Even in 1975, Calvet considered linguistic production as the production of individuals, having individual history as well as collective history, and experiencing linguistic and cultural exchanges which are affective, relying on power, concurrence and domination. For each individual, learning a language is not only a linguistic code to master (linguistics theories). It is a social act to master (sociolinguistics theories), a cognitive activity to develop (psycholinguistics theories) and intercultural processes to integrate (social psychology theories).

With so much emphasis attached to the symbolic power of language, language teaching and learning can no longer be considered as merely the

acquisition of linguistic, discursive and sociolinguistic competences. So far, linguistic proficiency, when referring to the sociolinguistic components, has put emphasis mostly on determined behavioural conventions, mismatching, misunderstandings in speech, turn-taking, expressions of politeness and some cultural points at the analytical level (*Common European Framework of Reference for Languages: Learning, teaching, assessment*, Council of Europe 2001). But, language acts are the medium of thought, culture and of representations that we have of other ethnic or religious groups and other cultures.

Assumptions about thought: how do we construct 'cultural representations'?

For Vygotsky (1962:51), language and thought are interconnected. Each word is a verbal act of thought and a 'microcosm' of the larger world. Moreover, verbal thought is determined by historical–cultural processes and language is the expression of such phenomena. Language is also the reflection of our experiences, our family, social and professional environment, our way of approaching our own culture and other cultures. As emphasised by Bourdieu (1982, 1994), language has an infinite capacity to generate relationships of symbolic values. Mental representations are constructed, that is schemes of perception and appreciation, of acquaintance and recognition, in which individuals invest their interests and their presuppositions. Thus, language acts are not pure and neutral linguistic elements. They must be conceptualised as cultural tools, vehicles of the culture and of the representations of other cultures that individuals make of the 'Other' and of other cultures. To understand the construct of mental representations, it is necessary to study social interaction, discourse, instructional and media discourse. No other semiotic code is as explicit as natural language in the expression of meanings, knowledge, opinions and various social beliefs in order to understand the role played by social actors (Van Dijk 1997).

In language education, cultural representations are omnipresent in the learning process. Learners, when confronted with real-life situations and cultural experiences, have to compose with their own cultural representations and those offered in the textbooks or introduced by the teacher or the host milieu.

Assumptions about culture: what is intercultural communicative competence?

UNESCO (2001:49) defines culture as 'a set of distinctive traits, spiritual and material, intellectual and affective which characterize a society or a social group'. They emerge through the interactions between individuals. It seems that to better understand another culture as a different culture it becomes important to take into consideration the discovery of the cultural specificity of the 'Other'. It becomes unthinkable to approach the cultural dimension without considering interculturalism.

Current research suggests that education should be the entry into culture; the 'culture of education' being seen as socialisation into cultural ways of knowing, believing, doing and valuing (Bruner 1996). Based on that assumption, to perform 'linguistically and culturally' well in another language or culture, an individual must learn about similarities and acceptance of differences within cultures while developing strategies to manage misunderstanding and confrontation. It becomes essential to transcend particularisms, to change negative behaviours and value positive attitudes towards others. According to Bourdieu (1994), as these changes take much time, working with adolescents lends itself to better results. By acting on their knowledge, skills and their existential knowledge, it becomes possible to influence their social world; the constituent power of a language being based on perception and thought.

A common framework of reference of ICC

One question remains: *What does the definition of culture in language teaching imply?* In a world of globalisation, we ask schools to educate students about citizenship and to make them aware of collective stakes. Such teaching should bring learners to reinterpret their own behaviours, attitudes and values, and confront them with those they meet in other cultures. Interacting effectively across cultures means accomplishing a negotiation between people based on both culture-specific and cultural-general features (Lussier, Golubina, Ivanus et al 2007). To that effect, teaching a second language may claim a significant role in educating future generations (Byram 1992). But, transforming knowledge into attitudes and values induces educators to broaden the learner's perceptual horizon. To reach such a goal, teaching intercultural communicative competence must be approached within logical coherence (Lussier 1997). Linear views of discursive practices cannot lead to a coherent and integrated epistemology of Otherness. The best example of such coherence was the importance of developing the *Common European Framework of Reference for Languages* (Council of Europe 2001) which is now the reference for language learning, teaching and assessment within the European context. The same framework has been developed for 'intercultural communicative competence' (Lussier 1997, in press) and was validated by the Canadian research (Lussier 2001, Lussier, Auger, Clément and Lebrun-Brossard, forthcoming). This framework is composed of three competences: knowledge, skills (know-how) and existential competence (being).

Domain of knowledge competence

Knowledge competence takes into consideration declarative knowledge. Three approaches are essential, each of them having the same relevance (Lussier 1997, in press). The *humanistic approach* refers to knowledge of the world related to collective memory and culture with capital 'C', the heritage

of civilisation. The *sociological approach* refers to knowledge linked to the socio-cultural context. It considers culture as a social phenomenon. It uses documents to generate facts, statistics and social, economical and political data. The *anthropological approach* refers to daily life situations or culture with a small 'c'. It is knowledge linked to the diversity in ways of living, habits and customs, verbal and non-verbal ways of behaving, institutions, norms, etc. It involves interpersonal relations and relations with members of other cultures and societies. It can also refer to the unspoken, beliefs, values and attitudes as specific to different cultures.

Domain of skills (know-how)

Skills are the use of knowledge in real life language situations. Three levels of skill are defined. The first level of competence allows the individual to function in the target language, 'linguistically' speaking. It refers to skills generally acquired in the classroom context as a learning process (Krashen 1981). At the second level of competence, the individuals need to experience plurilingual and pluricultural exchanges, out of the classroom in order to be able to adjust properly their behaviours to the social and cultural environment and to be able to interact efficiently. At the third level, individuals can master the language, linguistically and interculturally. They become able to decode messages which can carry different interpretations. They are able to negotiate conflicts and situations of misunderstanding. They can interrelate 'language, thought and culture' as key issues to the acquisition of intercultural communicative competence.

Domain of existential knowledge (being)

Existential knowledge relies on affective and psychological factors. It focuses on the development of attitudes and mental representations which shape our vision of the world. Three dimensions must be considered. The first dimension, known as *cultural awareness*, is well documented in the literature. It is a stage which implies the development of sensitivity to others and other cultures. It involves the understanding of similarities and differences among societies and cultures. The second dimension refers to a *critical appropriation or competence* of other cultures in connection with self-knowledge and self-identity. Individuals are in a position to adjust and adapt their ways of thinking towards their own culture and other cultures. It implies being able to accept and interpret other beliefs and to respect values associated with others and other cultures. The third dimension is known as *transcultural competence*. It implies 'the integration of new values, the respect of other values and the valorization of Otherness which derives from the coexistence of different ethnic groups and cultures evolving in a same society or in distinct societies while advocating the enrichment of identity of each culture in contact' (Lussier 2008).

Pertinence and impact of the conceptual framework of reference

Models of 'communicative competence' aiming at the development of linguistic, discursive and sociolinguistic competences are now considered to be incomplete. It is accepted that language, thought and culture are closely linked. Their interaction requires a new approach to the study of these concepts and their integration into the development of language curriculum, language teaching and assessment. They have to be revisited to include ICC.

To that effect, two studies using the conceptual framework can be consulted: 1) *Guidelines for the assessment of intercultural communicative competence* (Lussier, Golubina, Ivanus et al 2007), and 2) *Representations of others and other cultures in the context of the initial and ongoing training of teachers in European contexts* (Lussier, Auger, Urbanicova and Armengol 2003), both published by the Council of Europe and the European Centre for Modern Languages.

First study: Guidelines for the assessment of intercultural communicative competence

In language education, learners have to learn to interact with others and, consequently, to mediate between individuals of other cultures. This implies a negotiation between people and the criss-crossing of identities. It requires specific knowledge, skills and attitudes that are not taken into account in models of communicative competence. Thus, the need to evolve and to integrate ICC in the development of language curriculum.

As we know, focusing on the assessment of learning of knowledge would be easier. But, it becomes important to consider language teaching in terms of the appropriation of another culture, that is the development of cultural awareness, respect for other cultures, openness of oneself to diverse cultural experiences, etc. For that reason, there will always be some subjectivity in assessing ICC. The question is to look at ICC with the same understanding. The use of a common framework of reference is the answer to the problem. It guides the conception of assessment tasks, assessment procedures and methods to assess students' profile. In the study, educators can find examples for each of the ICC domains (knowledge, skills and existential competence). There are also examples of culture-logs, attitude questionnaires, self-evaluation, profile diagrams and the portfolio.

As important as language criteria is the development of criteria to assess ICC ability of the learner to perform a task. In order to reach common understanding, we need ICC benchmarks and competency levels which require 'rating on a scale'. That implies 'judging that a person is at a particular level or band on a scale made up of a number of such levels or bands' (Council of Europe 2001, North 1995). These scales already exist in language teaching. In the guidelines for the assessment of ICC, one of the major issues was

to come out with such scales and performance descriptors of the learner's general ability in the three domains of ICC. Here is an example, specific to the domain of 'existential knowledge' in terms of low, medium and high ability.

Table 1 Scale of ICC proficiency for 'existential knowledge'

Dimension	Low profile	Medium profile	High profile
Existential competence (being)	**Level of understanding** Cultural awareness Tolerance Sensitivity Realizing that there are differences in beliefs and values	**Level of accepting and interpreting** Critical appropriation Sympathy Openness Adjusting to different beliefs and values from other people and cultures	**Level of integrating and internalizing** Transcultural competence Empathy Sense of alterity Integrating new beliefs and values

With such profiles of levels of competence, researchers were able to define global performance descriptors which tell us what a learner can do to achieve each of the levels.

Table 2 Descriptors of ICC competence for 'existential competence'

Levels	Descriptors of ICC competence
Low	The student adopts a defensive approach to situations, and shows difficulties in intercultural experiences. The student expects adaptation from others, showing ethnocentric attitudes and perceptions. The student manifests tolerance to some culturally determined behaviors. The student applies cultural stereotypes and denotes a passive attitude towards other cultures.
Medium	The student manifests "mixed" attitudes to culturally determined behaviors. The student starts to accept intercultural ambiguities as challenging, showing openness and interest towards others. The student sometimes takes the initiative in adopting others' patterns; tends to see things and situations from the other's point of view. The student demonstrates openness to other cultures, accepting and being sympathetic to other beliefs and values. The student has no profound argumentation of his own position in terms of his own attitude regarding cultural differences.
High	The student enjoys observing, participating, describing, analyzing, and interpreting intercultural elements and situations. The student is well able to defend his own position toward different culturally determined acts or behaviors. The student expresses a sense of alterity, i.e. is able to reflect on what a person from a different culture would really feel like in such a given situation. The student expresses empathy toward representatives of different cultures. The student manifests respect for otherness, other beliefs and values. The student tries to take the role of a mediator in intercultural encounters, manages ambiguity, and offers advice and support to others, recognizing how one's world view is culturally conditioned.

A further step to specify these descriptors would be to define specific tasks in terms of behaviours and attitudes that teachers experience in the classroom. But, for the present study, there was no time for group discussions and interaction with teachers. On the other hand, there are levels of proficiency associated with examples of ICC tasks in the book.

Second study: Representations of others and other cultures in the context of the initial and ongoing training of teachers in European contexts

We know that teachers are social leaders as well as instructors. They convey cultural representations from various information sources: syllabus, teaching materials, texts used and their own experiences. But do they use such sources to develop intercultural communicative competence on the part of the learners? Do they see themselves as cultural mediators? Do they take into account identity processing? Do they adopt strategies to exploit, negotiate or even provide solutions when there are tensions or misunderstandings between groups of learners?

To answer these questions, a brainstorming session with researchers from different European countries was held where they were able to voice their expectations and needs in order to arrive at the following conceptual framework specific to the research.

Table 3 Conceptual framework on teachers' cultural representations and intercultural competences in the teaching of languages

Dimensions	Sub-dimensions		Number of items
1. Knowledge competence (Knowledge profile)	1.1 Education (Training)	1. In-Service	1–2
		2. On-going	3–6
		3. Linguistic	4
		4. Cultural	8
	1.2 Competences	1. Language L1–FL and MT	12 to 18–19 to 23
		2. Cultural	48–58–79
	1.3 Actual experience	1. Learning experience	5–7–30
		2. Linguistic and cultural practices	9 to 11–24–32 33 to 40
2. Skills (Behavioural profile)	2.1 Behaviour	1. Teaching approaches and methods	59 to 73 74–75
		2. Types of activities	41–76
		3. Strategies	28 to 31–77
	2.2 Mobility	1. Family/Social/ Professional	
3. Existential competence (Being) (Attitudinal profile)	3.1 Perceptions/ attitudes	1. Understanding	25–26–27
		2. Accepting/Interpreting	42 to 47–49 to 56
	3.2 Role as teachers	1. Tension /Negotiation	77–78
		2. Cultural mediation	57–80 to 84

This framework became the framework of reference for each of the countries involved in the research. Researchers were able to ensure the same orientation, the same aims and the same research questions by means of a survey questionnaire administered to a sample of European teachers. Such methodological procedure, based on common understanding, ensures the quality of the interpretations and leads to better decision-making or action in relation to aimed objectives.

Conclusion

Within the context of globalisation and the rapid expansion of plurilingual and multicultural societies, there is an increasing need for a common framework of reference of 'intercultural communicative competence'. The aim is to secure the links between the key concepts which determine and clarify a research domain, such as intercultural communicative competence. Such a framework allows the scientific community to duplicate research processes in terms of curriculum or target population and to transfer data in similar contexts or interrelated perspectives within similar characteristics. It gives curriculum and text designers guidelines to generate new programmes and textbooks. Finally, it gives teachers and evaluators new ways of approaching pedagogical practices, learning and assessment tasks.

References

Bourdieu, P (1982) *Ce que parler veut dire. L'économie des échanges linguistiques,* Paris: Fayard.

Bourdieu, P (1994) *Language and symbolic power,* Cambridge, MA: Harvard University Press.

Bruner, J (1996) *The culture of Education,* Cambridge, Ma: Harvard University Press.

Byram, M S (1992) Language and culture learning for European citizenship, in Beveridge, M C and Reddiford, G (Eds) {Special issue} *Language and education*, 6, 165–176.

Byram, M S (1997) *Teaching and assessing intercultural communicative competence,* Clevedon: Multilingual Matters.

Calvet, L J (1975) *Pour ou contre Saussure : Vers une linguistique sociale,* Paris: Fayard.

Calvet, L J (1999) *Pour une écologie des langues du monde,* Paris: PLON.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*, Modern Languages Division, Strasbourg: Cambridge University Press.

Galisson, R (1991) *De la langue à la culture par les mots,* Paris: CLÉ international

Kramsch, C (1998) *Language and culture,* Oxford: Oxford University Press.

Krashen, S D (1981) *Second Language Acquisition and Second Language Learning,* Oxford: Pergamon.

Lussier, D (in press) A conceptual framework of reference for the development of intercultural communicative competence.

Lussier, D, Auger, R, Clément, R and Lebrun-Brossard, M (forthcoming) *Validation of a conceptual framework of intercultural communicative competence: Empirical study of 'existential competence'*.

Lussier, D (1997) Domaine de référence pour l'évaluation de la compétence culturelle en langues, *Revue de didactologie des langues-cultures, ÉLA,* Paris: Didier Érudition 106, 232–246.

Lussier, D (2008) *Cultural representations and the construct of identity,* Ministry of Education of Quebec and Ontario: Research Seminar (unpublished).

Lussier, D, Auger, R, Clément, R and Lebrun-Brossard, M (2002–2008) *Cultural representations, ethnic identity and intercultural communication among young adults*, Ottawa: Social Sciences and Humanities Research Council of Canada.

Lussier, D, Auger, R, Urbanicova, V and Armengol, M (2003) Representations of others and other cultures in the context of the initial and ongoing training of teachers, in Zarate, G, Gohard-Radenkovic, A, Lussier, D and Penz, H (Eds) *Cultural mediation in language learning and teaching,* European Centre for Modern Languages, Strasbourg: Council of Europe Press. Chapter 7, 191–223.

Lussier, D, Golubina, K, Ivanus, D et al (2007) Guidelines for the assessment of intercultural communicative competence, in Lazar, I, Huber, M, Lussier, D, Matei, G and Peck, C (Eds) *Developing and assessing intercultural communicative competence: a guide for language teachers and teacher educators,* European Centre for Modern Languages, Strasbourg: Council of Europe Press.

Lussier, I (2001) *Validation du questionnaire «xénophilie/xénophilie» dans le contexte des représentations culturelles auprès d'adolescents montréalais,* MA Thesis, Université du Québec, Montréal.

North, B (1995) The development of a common framework scale of descriptors of language proficiency based on a theory of measurement, *System,* Great Britain: Pergamon, 23 (4), 445–465.

Sperber, D (1996) *Explaining culture, a naturalistic approach,* Oxford: Blackwell Publishers.

Stern, H H (1983) *Fundamental Concepts of Language Teaching*, Oxford: Oxford University Press.

UNESCO (2001) *Universal Declaration on Cultural Diversity*, Paris: UNESCO.

Van Dijk, T A (1997) *Discourse as Social Interaction,* London: Sage Publications.

Vygotsky, L S (1962) *Thought and Language*, Cambridge, Ma: The MIT Press.

14 The EIKEN Can-do List: improving feedback for an English proficiency test in Japan

Jamie Dunlea
The Society for Testing English Proficiency (STEP), Tokyo, Japan

Abstract

This paper describes the construction of the EIKEN Can-do List. The list was designed to help provide more detailed and useful feedback for a large-scale English proficiency test widely administered in Japan. The project utilised responses to self-assessment questionnaires from 20,000 test takers who had recently passed one of the seven levels tested by the EIKEN suite of English proficiency tests. The results obtained from the questionnaire survey were used to create a profile of what Japanese learners at these different levels of ability believe they can accomplish in English in real-life situations. The paper explains the procedures employed to create a can-do list specifically designed for Japanese EFL learners and discusses some of the methodological issues involved in the creation of the list.

Introduction

Background

As more and more individuals, along with institutions and governments, devote time and resources to English education, the importance of language testing in accurately and fairly assessing the ability of learners has also grown. To meet the changing needs of test users – learners, teachers, parents, and employers, among others – improving the interpretability and usefulness of language test results is a priority. In the past, reporting test results was often a matter of simply providing scores or some kind of level or grade classification, particularly with large-scale standardised tests. While this kind of feedback serves many important purposes, test users often request more

concrete information to help them understand what it means to have passed a certain level or attained a certain score in terms of using the language for communication outside the test or classroom. The EIKEN Can-do project was undertaken to help address this need, and specifically with the aim of increasing understanding of the levels of proficiency targeted by the EIKEN tests in Japan. The project utilised responses to self-assessment questionnaires from approximately 20,000 test takers who had recently passed one of the EIKEN tests to create a profile of what Japanese learners at different levels of ability, as defined by the seven-level EIKEN tests, believe they can accomplish in English in real-life situations. This paper describes the methodology and process of construction of the EIKEN Can-do List.

The EIKEN tests

The EIKEN tests are a seven-level suite of tests made and administered by the Society for Testing English Proficiency (STEP), a non-profit foundation established in Japan in 1963. As shown in Table 1, the seven levels of EIKEN are designated as 'grades', and range from Grade 5 (beginner) to Grade 1 (advanced), with two bridging levels (Grades Pre-1 and Pre-2). The Grades are designed to provide well-defined steps that can act as both motivational goals and concrete measures of English ability. Each grade is administered on a pass/fail basis. For Grades 3 through 1, the test is administered in two stages, with test takers who pass the first-stage written test required to sit and pass a face-to-face speaking test in order to achieve certification at that level. The tests are administered at sites across Japan and are taken by approximately 2.5 million test takers a year. This wide accessibility, combined with the large population of test takers, gives the EIKEN tests a unique potential to both contribute to and reflect the state of English language learning in Japan. The Ministry of Education, Culture, Sports Science and Technology (MEXT) has listed the EIKEN tests as benchmarks of recommended English

Table 1 Overview of EIKEN tests

EIKEN Grade	LEVEL	Stage 2 (Speaking)	Recognition/Uses
Grade 1	Advanced	Yes	International admissions to graduate and undergraduate programmes
Grade Pre-1		Yes	
Grade 2		Yes	MEXT benchmarks for high school graduates
Grade Pre-2		Yes	
Grade 3		Yes	MEXT benchmark for junior high school graduates
Grade 4		No	
Grade 5	Beginner	No	

ability for junior high school and high school graduates in the Action Plan to Cultivate Japanese with English Abilities (MEXT 2003).

The EIKEN Can-do List

Focus

The primary aim of the EIKEN Can-do List is to help test users gain a better understanding of the levels tested by the EIKEN tests. In its current form, the EIKEN Can-do List can be described as a 'user-oriented' scale, according to Alderson's three-way classification of user-oriented, assessor-oriented, and constructor-oriented scales (Alderson 1991). This is an important distinction and one which has influenced many of the methodological choices made in the process of construction. As it is designed to enable a wide variety of test users to better interpret the EIKEN levels, the structure of the list and the statements in it had to be clear and easy to understand, without overly technical language. This approach was the reason, for example, why the list was organised under the headings of the four skills – reading, listening, speaking, writing – rather than using other possible categorisation schemes. The four skills designations have wide currency in language learning in Japan and their use makes the categories in the list accessible to typical learners without the need for any extra explanation.

It is important to make a distinction between the focus of the EIKEN Can-do List and the content of the EIKEN tests themselves. The Can-do List looks more at the broad levels of English ability that the tests are designed to reflect, and attempts to describe what typical learners who have obtained a particular level are able to accomplish when using the language in real-world language-use situations. Meeting the necessary conditions to maintain language tests as highly reliable measurement instruments often precludes the inclusion of many truly unpredictable, real-world language use situations, particularly on tests administered on the scale of the EIKEN tests. The Can-do List utilised self-assessment methodology and so was able to investigate tasks that are difficult to include in the necessarily restricted format of the tests. The list also deals with some tasks and situations that are included on the EIKEN tests, and in so doing adds insight from the learners themselves about how they feel they would handle such tasks in the real world, outside the testing situation. The EIKEN Can-do List is thus one more important step in providing as much useful information as possible to test users, from as many different perspectives as possible. The results from the EIKEN tests and information about typical learner behaviours included on the EIKEN Can-do List are complementary. The two types of information add to our understanding of the EIKEN tests by helping test users understand more clearly what learners who have achieved the levels of language ability targeted by the tests are able to accomplish when using English.

Two subsidiary goals of the project are also relevant. Firstly, the list aims to contribute to a better understanding of typical language learners in Japan. Obviously, we need to be very careful about inferring too much information from the set of data on which the Can-do List was constructed, and we must be aware that its primary focus remains EIKEN test takers. But as mentioned above, the very wide availability of the EIKEN tests in Japan, and the very large sample size used when constructing the list – 20,000 test takers – mean that we can make some cautious inferences concerning typical Japanese learners based on the list. In this way, STEP hopes to provide information which will be of benefit to language educators and learners and thus make a positive contribution to language learning in Japan. The second subsidiary goal was to facilitate the comparison of EIKEN test takers and English language learners in the EFL context of Japan with test takers and learners in other contexts. By providing a highly reliable, empirically based snapshot of what EIKEN test takers believe they can do with English, STEP hopes to provide tools with which educators and researchers can achieve a better understanding of Japanese learners and compare the state of English language learning in Japan with international standards.

The EIKEN Can-do List was able to build on the ground-breaking work done for the creation of two major European projects: The Common Reference Levels that form the core of the Common European Framework of Reference for Languages (CEFR) and also the Association of Language Testers in Europe (ALTE) 'Can Do' project. These two projects have added a great deal to our knowledge of constructing descriptive scales of language proficiency, and in particular to the process of empirically scaling descriptors. Both of these projects were designed to be multi-lingual frameworks enabling the comparison of exams and qualifications across languages. The aims of the EIKEN Can-do List are more modest. While the EIKEN Can-do project took these important precedents into account, it has maintained its focus squarely on learners of English in Japan. The primary aim, as described above, is to provide insights into what EIKEN test takers, and by cautious inference, typical Japanese learners, believe they are able to accomplish in English. While individuals will always show some variation, perhaps not endorsing some descriptors that were calibrated for their level, or perhaps feeling confident at performing some tasks calibrated at a higher level, we can be confident that the list does in fact represent the degree of confidence of typical test takers who have passed the EIKEN tests.

Structure

The EIKEN Can-do List consists of 148 short can-do statements, also called descriptors, which describe the ability to use English to achieve various goals and complete real-life tasks. Statements in each of the four major skill areas

Table 2 Number of descriptors for each skill and grade

Grade	Reading	Listening	Speaking	Writing
1	4 descriptors	5 descriptors	4 descriptors	5 descriptors
Pre-1	5 descriptors	6 descriptors	7 descriptors	6 descriptors
2	6 descriptors	6 descriptors	6 descriptors	5 descriptors
Pre-2	4 descriptors	6 descriptors	6 descriptors	5 descriptors
3	5 descriptors	5 descriptors	7 descriptors	6 descriptors
4	6 descriptors	4 descriptors	4 descriptors	4 descriptors
5	5 descriptors	5 descriptors	6 descriptors	5 descriptors

**Each level of the list subsumes the can-do statements in each of the levels below it.*

– reading, writing, listening, speaking – are provided for each of the EIKEN grades (Appendix 1 provides the descriptors for Grade 3 as an example). An overview of the number of descriptors for each skill in each grade is provided in Table 2.

As mentioned in the previous sections, the EIKEN Can-do List is focused on learners of English in Japan rather than being a multi-lingual framework. It has also attempted to deal with some of the issues that Hasselgreen has identified as being important for young learners (2003, 2005). This focus has enabled the use of specific wording and examples which maximise the relevance and interpretability of the statements for Japanese learners of English. Two examples from Grade 5, one from Reading and one from Listening, will serve to illustrate this point.

> *Can recognize both upper case and lower case letters of the alphabet (e.g. A and a, F and f).*

> *Can understand which letter was referred to when the letters of the alphabet are spoken aloud (e.g. in the spelling of people's names).*

These descriptors extend the EIKEN Can-do List to learners at the very early stages of learning and particularly to young learners in formal education contexts. This is particularly important in an EFL context such as Japan where many learners only start to come into contact with the target language during formal education and many do not progress further than an elementary level of ability. The list recognises their stage of development and provides concrete examples of what other typical learners at their level can do, thus providing relevant, realistic yardsticks for these beginners. These two examples also take explicit account of a feature which is not exclusive to Japanese learners but certainly important – the difficulty of coming to terms with a completely different writing script from the scripts (kanji, hiragana, and katakana) which they use in everyday life.

Data collection

The use of self-assessment

The EIKEN Can-do project employed self-assessment questionnaires administered to test takers who had recently passed one of the EIKEN test levels. Self-assessment was chosen for several reasons. Firstly, the large pool of test takers for EIKEN provided access to a potentially huge sample of respondents. Utilising self-assessment allowed access to these numbers and this increased the robustness of the statistical analyses carried out. In total, data from over 20,000 respondents were utilised in the project. Self-assessment was the method of choice for the ALTE 'Can Do' project (Jones 2002), and a growing body of research provides support for the use of self-assessment in language learning and testing (Ross 1998, Blanche and Merino 1989). At the same time, while researchers have found a consistently significant correlation between self-assessment and criterion measures of language ability, we certainly need to be aware of the limitations of self-assessment. Ross (1998) in particular mentions the possible role of experience as a mediating factor in the reliability of self-assessment measures, and this is of course relevant to the EIKEN Can-do List. As we have already mentioned, the focus of the list is the EFL context of Japan. For many learners, particularly lower-level learners, there may be very few opportunities to actually use English outside their classrooms.

Questionnaires took account of this state of affairs and explicitly asked respondents to respond to all statements on the questionnaires, even if they did not have direct experience of such a situation. If they lacked experience, they were instructed to consider, to the best of their knowledge, how they believed they would handle such language use situations if the occasion arose. Despite the obvious drawbacks of this approach, it was felt that it provided the best opportunity to access the kind of information which the project aimed to gather: insights about real-life language use tasks and situations which could not be included in the controlled environments of language tests. While lack of experience may indeed affect learners' interpretations of the can-do statements and influence their judgments as to whether or not they could accomplish those tasks, it was felt that the potential shortcomings of self-assessment would be somewhat compensated for by the large number of respondents which self-assessment provided access to. The statistical procedures used to analyse the data allow us to say that at least the respondents were interpreting the statements and assessing their ability to handle them consistently, both within and across levels.

The CEFR employed an alternative to self-assessment, third-party teacher assessment of learners, during the major validation projects carried out for its construction (North and Schneider 1998, North 2000). North

(2000) provides a detailed description of the methodology involved, in which the teachers responded to questionnaires to judge their students' ability to handle the 'Can Do' statements on the questionnaires. For the EIKEN Can-do project, however, it was felt that third-party assessment by teachers would significantly reduce the potential sample size, while not necessarily providing better quality data. Third-party assessment definitely has the potential to provide valuable data on qualitative aspects of language use, particularly on aspects that learners may find difficult to judge, such as pronunciation or grammatical accuracy or complexity. However, lack of actual experience of the learners' performance in real-life language use situations is also a problem with third-party assessment. For example, the project to construct the CEFR found that some descriptors targeting content strands associated with work, such as using the telephone, meetings, and formal presentations, showed significant statistical misfit as teachers simply had no experience of their students' language use in these areas (North 2000:135). In Japan, where large classes, and particularly at high school, grammar-based, teacher fronted classes remain prevalent, it was felt that, in addition to the problem content strands identified in the CEFR project, teachers may indeed have little experience of judging their students' performance on many of the other real-life language use tasks described in the list. Given the potential problems inherent in either approach, it was decided to go as close to the source, the actual language user, as possible.

In the end, despite its drawbacks, self-assessment may in fact be the only truly appropriate way to gain insight into the areas of real-life language use which the EIKEN Can-do project was targeting, and certainly the best way to do so on the large scale necessary for the kind of analyses that were carried out. The Can-do List in its present form, it should also be remembered, is not intended to be used in place of tests. It is an additional, potentially powerful source of information which should be used not only in addition to language test results but where possible in conjunction with other relevant information obtained about and from test takers and teachers. Certainly, STEP recognises the need for future research to add various perspectives. Combining self-assessment with third-party assessment, assessing learners' performance on test tasks that operationalise various can-do statements, and investigating the link between respondents' actual experience of the situations in the can-do statements and their willingness to endorse those statements are some of the future projects being considered. While potentially offering new and important information, all of these methods will entail smaller sample sizes with coverage of a more limited range of content than was obtained through self assessment. In fact, during the construction of the current list, a certain degree of triangulation was attained to bolster the self-assessment data by administering a supplementary survey to 8,000 test takers who also responded to the Can-do questionnaires. This English Use and Study Habits

Survey (STEP 2006) asked questions such as how much test takers actually used English outside the classroom in their daily lives and what kinds of English-language materials they used inside and outside the classroom, such as the internet, newspapers, and movies. This information was used, for example, when investigating statements that did not perform as expected.

Writing descriptors

A pool of descriptors was written by a project team of staff members involved in the production and analysis of the EIKEN tests. In writing the descriptors, reference was made to a number of sources, including the CEFR and ALTE 'Can Do' projects mentioned earlier, as well as others such as the DIALANG self-assessment lists, which themselves are based on the CEFR, and the ACTFL Proficiency guidelines. A number of Japan-specific sources were also consulted, including the Courses of Study curriculum guidelines for lower and upper secondary school (MEXT 2003a, 2003b), textbooks and learning materials commonly used in formal education as well as in language schools, and TV and radio language education programmes. The reference to these learning materials was important for two reasons. While the aim of the EIKEN Can-do List is, as stated above, to bridge the gap between language tests and language use outside testing environments, it is also true that for many learners in an EFL context the classroom will constitute a significant part of the target language use domain. In their description of the target language use domain (TLU), Bachman and Palmer (1996) take account of this reality and describe two general types of TLU, real-life domains and a language instruction domain, which they also note are not always completely distinct. Several descriptors take explicit note of using English for study and training, and some, such as the following, provide examples which link the descriptor to situations typical of those in which learners use English to study English, either in Japan or in language schools overseas. While classroom based, the descriptor below is nonetheless a case of using English to achieve a real-world purpose – learning a language or improving specific language abilities.

> *Can understand classes and training courses conducted in English, provided that the content is simple (e.g. foreign cultures and lifestyles in foreign countries).*

The second reason for consulting these learning materials was to be aware of the meta-language that learners would be familiar with when describing aspects of language use, and also to be aware of the kinds of language activities and language practice tasks they would be familiar with. This was very important because of the user-oriented nature of the list. As noted above, the

list is designed to be accessible to a wide range of test users, and so the language needs to be easy to understand and accessible, using everyday language rather than specialised jargon which might be familiar to language educators but not necessarily to language learners. Assuring the ease and consistency of interpretability of the can-do statements by learners of different ages and abilities was also an essential part of obtaining high-quality data from the wide range of test takers who took part in the survey. By taking note of the language learning activities and tasks with which learners were familiar, it was also hoped that the problem of lack of experience could be mitigated to some extent. When writing descriptors for which the relevant test takers may not have had experience, the descriptors were worded to allow test takers to make reference to common tasks that they may have had experience of in a classroom situation. This, it was felt, would provide some basis of experience for them to refer to when considering how they would handle such tasks in the real world.

As far as possible, the guidelines noted by North (2000) for making quality descriptors of language proficiency were followed: positiveness, definiteness, clarity, brevity, independence. Even when following these guidelines, however, as Morrow (2004) notes, some ambiguity of interpretation will remain. Morrow cites the following B1 CEFR descriptor and comments that terms such as 'main points' and 'clear' will potentially have different meanings for different people.

> *Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.*

For this reason, it was decided to make judicious use of examples in the EIKEN Can-do descriptors to illustrate the kind of activity or the quality of performance that was expected. This was especially important for EIKEN, which has a wide and diverse demographic profile of test takers. Examples not only helped to fulfil the list's purpose of being user oriented, by communicating clearly and simply what descriptors were trying to say in a general sense, but they also helped to maintain consistency in interpretation across different groups of test takers by clearly indicating the context of language use and/or the degree of ability intended by individual statements.

The tasks included in the EIKEN tests were also an invaluable resource in designing and refining the can-do statements. As Bachman and Palmer (1996:60) note, for the last several decades 'language teaching methodology. . . has been aimed at creating teaching and learning tasks in which the purpose for using language is to communicate, and not simply to learn'. Language testing has of course reflected these changes, and the EIKEN tests are continuously reviewed, and when appropriate revised, to try to ensure that the test content contains communicative tasks that are meaningful and

useful for language learners. As such, the EIKEN Can-do List makes reference to many kinds of language-use tasks that are reflected in the test tasks in the EIKEN tests themselves.

After an initial pool of descriptors was created, they were then provisionally assigned to different grades. To do this, reference was made to the level of difficulty and sophistication of tasks in the EIKEN tests. The data accumulated on the difficulty of tasks administered over 50 years to more than 76 million test takers by STEP thus proved a valuable resource for comparing and assessing the potential difficulty of the can-do statements, and this process often led to textual revisions to enhance the appropriacy and clarity of particular descriptors for particular levels. This process allowed for extremely accurate targeting of descriptors, with content and examples matched closely to the appropriate levels.

The questionnaire format

Each questionnaire contained approximately 70 to 100 descriptors, depending on the level targeted. All of the questionnaires were administered in the test takers' native language, Japanese. Each descriptor was accompanied by a five-step, Likert-type scale representing various degrees of confidence to perform that particular task. An example of a descriptor and the five-step scale is included below.

> *Can write simple cards and postcards (e.g. birthday cards, postcards sent while on vacation).*
> 1 2 3 4 5

The five-step scale was explained on the first page of the questionnaires, and the following brief descriptions (in Japanese) of the scale were repeated at the top of each page of the questionnaires.

> *1 (I think) I cannot do this*
> *2 (I think) I can only do this to a very limited degree*
> *3 (I think) I can do this to some degree*
> *4 (I think) I can generally do this*
> *5 (I think) I can do this very well / with ease*

Subjects

Questionnaires were administered over three rounds, as shown in Table 3. In each round, 2,000 test takers from each grade were randomly selected from those who had passed the most recent administration of the EIKEN tests. The samples were controlled to provide a similar demographic profile

Table 3 Number of respondents in each round of questionnaires

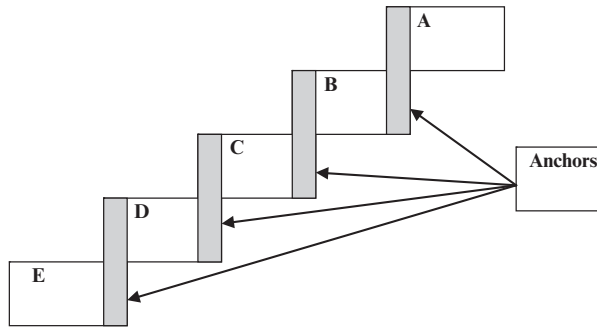
GRADE	1st Round Dec, 2003	2nd Round Feb, 2004	3rd Round Dec, 2004	TOTAL
1	1,200	—	1,267	2,467
Pre-1	1,247	—	1,320	2,567
2	1,193	1,260	1,277	3,730
Pre-2	1,069	1,249	1,123	3,441
3	1,070	1,150	1,148	3,368
4	1,020	1,097	1,078	3,195
5	989	—	972	1,961
TOTAL	7,788	4,756	8,185	20,729

to that of the test takers for the actual EIKEN tests, controlling for sex, age, and passing scores. A smaller pilot administration was also carried out in the summer of 2003 to assess the appropriacy of the proposed questionnaire and the five-level response format. The response rate for all grades in all three rounds was consistently high, staying above 50% for all grades except Grade 5. During the second round, only three sets of questionnaires were used, as these were targeted at the mid-levels of the EIKEN tests to help to flesh out perceived gaps in the numbers of descriptors that had been adequately calibrated to these levels during the first round.

Linking questionnaires

For each round, several questionnaires were produced – five each for the 1st and 3rd rounds, and three for the 2nd round. Each questionnaire was linked to the questionnaires above and/or below it by common anchor items, as shown in Figure 1. In this way, Questionnaire B was linked to A by anchor items shared between A and B, and B was also linked to C by a different set of anchor items shared between B and C, and so on.

Figure 1 Anchoring questionnaires



Each questionnaire was targeted at different levels and administered to test takers who had passed the appropriate grades. Table 4 shows the grades that each questionnaire was administered to and the grades that the descriptors it contained were targeted at. In this way, each questionnaire contained descriptors closely matched to the level of the respondents and some descriptors slightly above and/or below that level.

Table 4 Level of subjects and intended level of descriptors for questionnaire forms

Questionnaire	Subjects	Intended level of descriptors contained in questionnaire
A	Grades 1 & Pre-1	*1 Pre-1 2*
B	Grade 2	*Pre-1 2 Pre-2*
C	Grade Pre-2	*2 Pre-2 3*
D	Grade 3	*Pre-2 3 4*
E	Grades 4 & 5	*3 4 5*

Data analysis

Placing descriptors into a hierarchy of difficulty

Data from the three rounds of questionnaires were analysed using the graded response model, a polytomous IRT model. The graded response model was chosen as we were particularly interested in finding out how the degree of confidence of test takers changed according to the quality of performance expressed in the five-point response scale, and this model is particularly suited to dealing with polytomous response data (Saida 2007, Tang 1996). Following each round, results were analysed and items that had performed differently from expected were dropped, reassigned to new levels, or rewritten to avoid suspected problems of interpretation. The final data set, then, included data from many descriptors which had been administered in all three rounds, providing those descriptors had shown sufficient stability across administrations. When descriptors displayed problems and were rewritten, only results from the edited versions were included in the final analysis. The fourth step on the response scale, phrased as *(I think) I can generally do this,* was eventually chosen as the criterion for the confidence level of participants, and all of the descriptors were then placed on a common scale using this large data set. Descriptors within each of the four skill areas were then placed into a hierarchy of difficulty based on their standing on the scale relative to descriptors within the same skill area. The results of this analysis were treated as the primary condition for confirming whether the descriptors actually belonged in the grades they had been provisionally assigned to. Where

descriptors were calibrated at a level higher or lower than intended, the first step was to go back and look at the content of the descriptor and investigate why test takers had responded to them in this way. During this process, reference was made to the English Use and Study Habits survey mentioned above, as well as other sources such as test results and demographic information. It is significant that all respondents to the EIKEN Can-do project had been tested and placed at one of the EIKEN levels. Thus not only a rich array of test data but also demographic information were available for every respondent.

Confirming the relationship with EIKEN grades

As noted above, all of the descriptors had been provisionally assigned to levels based on an extensive review of EIKEN test data and correspondence to tasks and activities typically found in learning materials commonly used in Japan. The empirical analysis was used to confirm whether these author intentions, as North (2000) refers to them when discussing the setting of cut-offs, were supported by the data. The results of the IRT analysis described above demonstrated that descriptors performed on the whole as expected: descriptors within each skill area provisionally assigned to Grade 5 were rated lower than descriptors within the same skill area assigned to Grade 4, Grade 4 descriptors were rated lower than those in the Grade 3 group and so on. Where the difficulty level of individual descriptors placed them above or below the majority of other descriptors in the same skill area for the grade for which they were intended, the procedures described in the section above were followed to decide whether to drop the descriptor or to put that descriptor into the grade corresponding to its difficulty level. The hierarchy of difficulty that was obtained by the IRT analysis was thus given precedence over author intentions in determining levels, but the overall coherence of the authors' intentions was in fact confirmed by the way the majority of individual descriptors were grouped into a hierarchy of difficulty within each skill area corresponding to the hierarchy of the grade levels for which they were intended.

To then confirm the intended relationship with EIKEN grades, an extra set of conditions was imposed relating to the response rate of test takers endorsing the different levels of confidence on the five-step scale. As noted above, an important feature of the EIKEN Can-do Project was the way in which questionnaire forms were able to be tailored to the ability level of respondents, which was known in advance as all respondents had recently passed one of the levels of the EIKEN tests. As shown in Table 4, respondents who had recently passed the Grade 2 level, for example, were sent a form containing all the descriptors intended for Grade 2, as well as descriptors designed for the levels immediately above and below. For a descriptor to be confirmed as belonging to a certain grade, 80% of respondents at that level

had to have chosen 3 or above, and 50% had to have chosen 4 or above on the five-step response scale. This was done to confirm, for example, that not only were reading descriptors intended for Grade 5 easier than those intended for Grade 4 based on their position on the scale, but that according to the criteria described above, a majority of respondents who had passed the Grade 5 test did in fact endorse the reading descriptors intended for Grade 5.

Once the empirical analysis had established the hierarchy of descriptors and confirmed the specific grades to which the descriptors belonged, the entire list was reviewed to confirm the coherency of the content. At this stage, several descriptors were dropped despite the fact that they had shown sufficient stability in the empirical analysis. This occurred when the content of these descriptors was thought to overlap to some degree with other descriptors calibrated to the same level. As they did not add new information, it was felt these descriptors would cause some confusion to users of the list, and so it was decided to drop these descriptors from the final list. These descriptors, along with those that were dropped as result of the various statistical analyses employed, are being held in a reserve descriptor bank for future reference, and may prove useful as a starting point for writing new descriptors in future efforts to expand the width and depth of coverage of the present list.

Conclusion

The EIKEN Can-do List is by no means intended to be exhaustive. The EIKEN Can-do project has provided a robust, solid foundation of a core of general can-do statements which describe a relatively small range of tasks chosen because of their relevance to as many learners in the Japanese EFL context as possible. This was important to ensure they would be interpreted consistently by EIKEN test takers, who come from a wide range of ages and profiles. This focus on clarity, coupled with the unprecedented amount of information, including test scores, available to STEP, allowed descriptors to be very closely targeted to test takers at an appropriate level, and this in turn ensured high quality data.

Future research will focus on increasing both the depth and breadth of the list, by, for example, focusing on particular content strands, such as shopping or taking part in lectures or meetings, and particular target groups, such as university students, etc. At this stage, descriptors for such areas are often available as single statements for the level for which they were thought most relevant. Future research aims to not only increase the breadth of the list by increasing the number and range of tasks covered, but also to increase the depth by targeting these single descriptors and expanding them to establish different degrees of performance across different levels. It is also worth re-emphasising that the EIKEN Can-do List is a valuable tool to be used in conjunction with other information, including test scores, to further our

understanding of EIKEN test takers and Japanese English language learners. It is not intended to be able to answer all of our questions or fulfil all of our needs regarding test takers' abilities. And, indeed, in language education and testing in particular, no such one-size-fits-all tool or method probably exists. Provided we remember these caveats, the list has the potential to be a powerful source of information, and can be said to provide the most comprehensive snapshot currently available of what EIKEN test takers, and by extension Japanese learners, can accomplish in English in real-life language-use situations.

Acknowledgements

The EIKEN Can-do project was a collaborative effort involving the cooperation of many individuals. In addition to the author, the project team consisted of: Kazutaka Tanimoto, Shoji Akai, Kazuaki Yanase, Nami Sasaki, Tomoki Matsudaira, and Yumi Uno. Each member devoted many hours of hard work to the process of constructing the list. Special mention should be given to Kazutaka Tanimoto, the project coordinator, Kazuaki Yanase, who wrote original drafts of many of the descriptors for the team to review, and Tomoki Matsudaira, who oversaw the data analysis.

References

Alderson, J C (1991) Bands and scores, in Alderson, J C and North, B (Eds) *Language Testing in the 1990s*, London: Macmillan.

ALTE (2002) *The ALTE CAN DO PROJECT (1992–2002)*, <www.alte.org/can_do/alte_cando.pdf>.

American Council for the Teaching of Foreign Languages (1999) *ACTFL Proficiency Guidelines Speaking*, <www.actfl.org/files/public/Guidelinesspeak.pdf>.

American Council for the Teaching of Foreign Languages (2001) *ACTFL Proficiency Guidelines Writing*, <www.actfl.org/files/public/writingguidelines.pdf>.

Bachman, L F (1991) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

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

Blanche, P and Merino, B (1989) Self-assessment of foreign language skills: implications for teachers and researchers, *Language Learning* 39 (3) 313–340.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*, Cambridge: Cambridge University Press.

Hasselgreen, A (2003) *Bergen Can Do Project*, Council of Europe Publishing.

Hasselgreen, A (2005) Assessing the language of young learners, *Language Testing* 22 (3) 337–354.

Jones, N (2002) Relating the ALTE framework to the Common European Framework of Reference, in *Common European Framework of Reference*

for Languages: Learning, Teaching, Assessment. Case Studies, Strasbourg, Council of Europe Publishing.
Ministry of Science, Education, and Technology (2003) *Regarding the Establishment of an Action Plan to Cultivate "Japanese with English Abilities"*. Retrieved September 2, 2005, from <www.mext.go.jp/english/topics/03072801.htm>.
Ministry of Science, Education, and Technology (2003a) *The Course of Study for Lower Secondary School: Foreign Languages*. Retrieved September 2, 2005, from <www.mext.go.jp/english/shotou/030301.htm>.
Ministry of Science, Education, and Technology (2003b) *The Course of Study for Upper Secondary School: Foreign Languages.* Retrieved September 2, 2005, from <www.mext.go.jp/english/shotou/030301.htm>.
Morrow, K (2004) Background to the CEF, in Morrow, K (Ed.) *Insights from the Common European Framework*, Oxford: Oxford University Press.
North, B (2000) *The Development of a Common Framework Scale of Language Proficiency*, New York: Peter Lang Publishing, Inc.
North, B and Schneider, G (1998) Scaling descriptors for language proficiency scales, *Language Testing* 15 (2) 217–262.
Ross, S J (1998) Self-assessment in second language testing: a meta-analysis and analysis of experiential factors, *Language Testing* 15 (1) 1–20.
Saida, C (2007) Comparison of Concurrent Calibration of English Achievement Tests under Dichotomous and Polytomous Item Response Models, *Japan Language Testing Association Journal* 10, 119–133.
Society for Testing English Proficiency (STEP) (2006) 英検合格者の英語学習使用調査の報告 [Report on the English use and study habits survey of test takers who have passed the EIKEN tests]. STEP英語情報7・8月号, pp. 25–29.
Tang, K (1996) *Polytomous Item Response Theory Models and Their Applications in Large-Scale Testing Programs: Review of Literature*, New Jersey: Educational Testing Service.

Appendix 1
Can-do statements EIKEN Grade 3

The English translation of the full EIKEN Can-do List can be downloaded from the STEP website at the following address: www.eiken.or.jp/about/cando/Eiken_CandoList_translation.pdf

Grade 3

Speaking
Can take part in simple interaction about familiar things and talk about himself/herself.
Can talk briefly about something that he/she is interested in (e.g. his/her hobbies, club activities).
Can say what he/she likes and dislikes and explain in simple terms why (e.g. animals, food, sports).
Can describe routine actions from everyday life (e.g. "I got up at seven." / "I ate some bread for breakfast.").
Can describe simple plans (e.g. "I'm going to meet my friends.").
Can make simple requests (e.g. "Can you open the window, please?").
Can make invitations to familiar places and events (e.g. "Let's go to a movie tonight.").
Can use simple fillers, interjections, and responses in conversation (e.g. "I see." / "Really?").
Writing
Can write simple texts about himself/herself.
Can write a simple self-introduction (e.g. name, where he/she lives, family).
Can write about his/her hobbies or interests.
Can write what he/she likes and dislikes and explain why (e.g. food, sports, music).
Can write a short diary entry (from one to three sentences).
Can write simple cards and postcards (e.g. birthday cards, postcards sent while on vacation).
Can write short messages (e.g. "Ken called at 3 p.m.").
Reading
Can understand simple stories and texts about familiar things.
Can understand simple texts about topics that he/she is interested in.
Can understand texts about familiar topics related to everyday life (e.g. sports, music).
Can understand short, simple stories (e.g. simple biographies, fairytales).
Can read simple reading materials that include footnotes and explanations in Japanese (e.g. school reading materials, stories written for learners).
Can find streets, shops, and hospitals, etc., on simple maps written in English.

Grade 3 (continued)

Listening
Can understand instructions and talks and monologues about familiar things, provided that the speaker speaks slowly.
Can understand talks and monologues about topics that he/she is interested in, provided that the speaker speaks slowly and/or repeats sections (e.g. things related to his/her hobbies, music and sports that he/she likes).
Can understand the content of simple talks and monologues about familiar topics related to everyday life, provided that the speaker speaks slowly and/or repeats sections (e.g. school, club, activities, talking about the weekend).
Can understand simple announcements, provided that the speaker speaks slowly and/or repeats sections (e.g. meeting place, arrival and departure times for transportation).
Can understand simple directions, provided that the speaker speaks slowly and/or repeats sections (e.g. "Go straight and turn left at the next corner.").
Can understand words that are linked when pronounced in connected speech, provided that they are commonly used expressions (e.g. "Come in." as "C'min." / "Don't you" as "Doncha?").

* The statements in bold at the head of each skill section are summaries of the significant features of the can-do statements calibrated to that level for that skill. The summary statements themselves were not included in the questionnaire surveys.
* Each level of the EIKEN Can-do List can be assumed to include the can-do statements in each of the levels below it.

15 Democratising and enhancing the quality of institutionalised language assessment through the European Language Portfolio

Stergiani Kostopoulou
Trinity College, Dublin

Abstract

Taking account of concerns in the field of language assessment, this paper argues that integrating learner self-assessment based on the European Language Portfolio (ELP) into institutionalised assessment could result in a model of language assessment that is 'more educational, democratic, ethical and, at the same time, valid' (Shohamy 2001:390). A practical example of implementation is offered from English language support classrooms for immigrant students in Irish post-primary schools to demonstrate the pedagogical and social value of ELP assessment for immigrants learning the language of the host country.

The aims and scope of the current language assessment culture

The field of language assessment and testing has undergone major reforms in terms of its general scope, aims and functions, content and methods due to advances in the domains that inform it (Applied Linguistics, Second Language Acquisition research, educational theory, validation theory etc.). New conceptualisations of the construct of language proficiency, insights into learners' developmental patterns in learning and studies of the washback effect of assessment are some of the factors that point to the need to bring language learning, teaching and assessment into a closer interaction than has usually been the case. The Common European Framework of Reference for Languages or CEFR (Council of Europe 2001) constitutes an attempt

towards this direction, as its sub-title 'Learning, teaching, assessment' implies, by providing 'a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe' (Council of Europe 2001:1).

At the same time, the recognition of the social dimension of the field, on the grounds that assessment focuses on social aspects of language use while it simultaneously has a powerful impact on society (e.g. Boud 1995, Broadfoot 2001, McNamara 2001, McNamara and Roever 2006, Shohamy 1998, 2001), has led to a re-examination of the priorities and responsibilities of researchers. Establishing ethics and standards of practice and democratising assessment to prevent abuses of its power (Shohamy 2001) are major current concerns. These concerns become even more critical in the case of assessment practices for immigrants learning the language of the host country. Questions such as the one posed by Cummins (1997) reflect the issues of power and control that are inherent in assessment policies and in educational structures in general: 'Are we preparing students to accept the societal status quo (and in many cases their own inferior status therein) or are we preparing them to participate actively and critically in their society as equal partners with those who come from dominant group backgrounds?'. They also respond to McNamara's (2008) appeal to language testers to be thinkers and not just technicians.

All of the above indicate that the agenda of the current assessment culture should focus on enhancing the quality of language assessment and testing and strengthening the links between learning, teaching and assessment while ensuring that assessment practices and procedures are democratic and ethical and do not violate learners' rights and particularly the rights of diverse learners in multicultural societies (McNamara and Shohamy 2008). From a broader perspective, language assessment should also promote learners' holistic growth and learner autonomy which can be identified as two principal aims of second/foreign language (L2) education, given the shift to more learner-centred and humanistic pedagogies and the need for lifelong learning in the knowledge society.

Learner self-assessment through the ELP is a form of assessment that could substantially contribute to the above if it is appropriately integrated into existing assessment procedures in formal education.

The ELP as an instrument for learner self-assessment

The ELP is a language learning and reporting instrument developed by the Language Policy Division of the Council of Europe to promote its key political aims: recognising linguistic and cultural diversity, promoting tolerance and the development of plurilingualism and supporting education for democratic citizenship. It consists of:

a) a *language passport* in which learners self-assess and summarise their linguistic identity, their language learning and language qualifications in an internationally transparent manner,
b) a *language biography* which enables learners to assess themselves, set learning targets, monitor their progress on the basis of functional ('I can') checklists and record learning and intercultural experiences, and
c) a *dossier* in which learners keep samples of work that best represent their L2 proficiency. (*Council of Europe 2000 : 3*)

This tripartite structure of the ELP enables learners to document and report their proficiency in different languages in a comprehensive and transparent way to inform external educational authorities, future employers etc. Additionally and more crucially, it stimulates pedagogical processes that foster learners' intrinsic motivation and the capacity for language learner autonomy by encouraging them to set goals, reflect on and assess their learning progress, and gradually become responsible for their own learning. The reporting and pedagogical functions of the ELP are, in essence, realised through learner self-assessment, which is central to effective ELP use.

Learner self-assessment in the language passport, biography and dossier

Language learners are engaged in a constant process of formative and summative self-assessment in the three components of the ELP on the basis of the common reference levels of the CEFR (Council of Europe 2001). The language passport encourages a summative form of self-assessment as learners focus on the outcomes of the L2 learning process in order to provide an overview of their language proficiency, according to six levels (A1, A2, B1, B2, C1, C2) and five skills (listening, reading, spoken interaction, spoken production, writing), at a certain time using the scales and descriptors of the CEFR.

In the language biography, learners are invited to assess their learning progress according to functional ('I can') checklists arranged by levels and skills on a regular basis. This component requires learners to reflect on and assess the 'process aspect' of learning which implies formative self-assessment. This becomes, as Little and Perclovà (2001:55) argue, 'as much a habit of mind as an activity' because it forms an integral part of the language learning experience. These 'I can' checklists have multiple functions because they not only provide assessment criteria for self-assessment and assessment by external agents but they also reflect teaching objectives, communicative tasks and enable learners to set learning goals, monitor their progress, identify possible areas of weakness and plan further learning. In this respect, self-assessment in the ELP not only instigates a process of evaluation of what has been achieved but also motivates learners to think ahead and plan future action.

The use of the dossier also requires learners to think critically in order to select samples of work that correspond to their attestations made in the passport and biography and to regularly update these as their proficiency level changes. Maintaining their dossier effectively presupposes that they are capable of assessing their progress and learning outcomes in order to select the appropriate materials as evidence of their learning achievements.

According to the Council of Europe's *Principles and Guidelines* the ELP belongs to learners (Council of Europe 2006) and they are, therefore, responsible for it, as a physical object, but also for the pedagogical procedures it stimulates including assessment. For this reason, 'teacher assessment should always be separate [from the learner's self-assessment] and not used to correct it' (Council of Europe 2006:11). It is, however, pointed out that 'separate spaces for assessment by others must be available elsewhere' (Council of Europe 2006:5). Therefore, self-assessment should be separate from but scaffolded by and combined with teacher assessment. In this way, learners' responsibility for self-assessment is respected and teachers' mediating role in helping learners develop self-assessment skills is underlined. Combining learner self-assessment based on the ELP with external assessment is also necessitated by the challenges and limitations that are inherent in the practice of learner self-assessment of language achievement and proficiency.

Issues of feasibility, validity and reliability

Self-assessment in the ELP focuses on language learning for functional purposes through the 'I can' checklists, reflecting in this way the CEFR's action-oriented approach. These checklists facilitate self-assessment as learners at all levels might easily identify the communicative tasks they can perform in the target language (TL) and to the extent they can do so, and assess their behavioural capacity against the 'Can Do' statements. Forming judgements about the qualitative aspects of their language use (e.g. grammatical accuracy, phonological control, sociolinguistic appropriateness), however, is very challenging for learners (Little 2005:327) as these lie at least partly beyond the scope of introspective self-assessment in language learning. Little also points out that 'at the lower levels (A1 and A2) self-assessment against the checklists is relatively straightforward because the descriptors mostly refer to simple tasks; whereas the further you go up the levels, the more complex the descriptors become, so that self-assessment involves a lot of detailed analysis of the tasks captured in the descriptors' (personal communication).

Complexities entailed in self-assessment such as the above give rise to questions about learners' ability to undertake self-assessment effectively. Further, the very nature of self-assessment inevitably raises doubts about the subjectivity and accuracy of learners' judgements. These issues of feasibility, validity and reliability are crucial for the quality of learner self-assessment in

the ELP as they are for any other kind of assessment. With regard to the feasibility of self-assessment, there is a growing body of research data in support of learners' ability to produce accurate and reliable judgements of language proficiency (e.g. Bachman and Palmer 1989, Blanche 1990, LeBlanc and Painchaud 1985, Oskarsson 1978, 1984, 1989, Ross 1998). By eliciting learners' self-assessments through various instruments (e.g. questionnaires, multilevel rating scales etc.) and assessing these against external criteria (written test scores and teacher evaluation), these studies resulted in high correlations between self-assessment and teacher assessment (cf. Oscarson 1997). This suggests that, like any other skills, self-assessment skills can be developed through practice and learner training – both technical and psychological (Brindley 1989, Dickinson 1989, Holec 1985, Oscarson 1997). Kohonen (2002:87) suggests that learners can be introduced to the practice of self-assessment by developing a 'basic reflective orientation' in relation to themselves as language learners first and subsequently focus on their language proficiency.

Concerning validity and reliability, it should be emphasised that the criteria that apply to language testing cannot apply to self-assessment based on the ELP. This is because the latter falls outside the paradigm of traditional assessment as it is based on a different philosophy of learning and addresses different educational goals. ELP self-assessment is underpinned by theories of learner-centredness and learner autonomy and it is not based on the tradition of psychometrics. Learners – and not teachers – are responsible for forming judgements and the focus of assessment extends beyond the measurement-driven scope of language testing to encompass aspects of the learning process as well. These facts lead to a different conceptualisation of validity and reliability which, in the case of self-assessment, refers to 'the extent to which learners' estimates of their own language skills are consonant with independent ''objective'' criteria such as test results and teacher evaluation' (Oskarsson 1984:31). It is important that learners 'have at their disposal a measuring standard by which they may express their intuitions' (Oscarson 1997:178) and assessment criteria which are achievable, specific and relevant to their personal learning needs.

In the ELP, the 'I can' checklists (e.g. I can understand important school rules, I can give a short talk about the country I come from and explain about my culture and beliefs etc.) provide criteria which are 'achievable, specific and relevant to [learners'] personal learning needs' and teacher assessment and language tests that are used in combination with ELP assessment offer the 'independent ''objective'' criteria' that validate and complement learners' judgements. Providing evidence for learning achievements in the dossier is also fundamental for considering learners' claims as valid. As far as the issue of subjectivity is concerned, Oskarsson remarks that it 'does not necessarily invalidate the practice of self-assessment techniques' and that 'self-assessment may be motivated by reasons that go beyond mere evaluation' (1989:2). It may be that it fulfils the learners' need to regulate their own

learning and generally be the 'origin' of their actions, a state which is key to intrinsic motivation (deCharms 1968, cited in Deci 1996:27).

But why should ELP self-assessment be included in a language assessment culture? Allowing for its afore-mentioned functions and the processes it stimulates, self-assessment based on the ELP should be incorporated into institutionalised assessment because it offers an additional approach to forming judgements about learners' L2 achievement and proficiency. These judgements are based on evidence which is collected by learners themselves in the dossier. ELP self-assessment is a manifestation of assessment as a constructive instrument for learning apart from a tool for measuring achievement (as it concerns both the content and process of learning). It also reveals certain affective and psychological factors and aspects of personal and social development of learners that impinge on their observable performance and are inaccessible to external assessment (Brown 1990:9, Satterly 1989; for instance learners' ability to evaluate the effort they make, how hard they work, how far they think they have achieved their learning objectives etc.). As such, self-assessment complements and is complemented by teacher assessment, language testing and external assessment and their combination can result in a more balanced and holistic form of assessment.

But in addition to enriching and enhancing the quality of the existing assessment procedures and the results that are reported, ELP self-assessment can have a very positive impact on learners' holistic growth, as it is discussed below.

Integrating ELP assessment into institutionalised assessment: a 'creative' tension

Both summative and formative self-assessment through the ELP can be deployed to guide the management and assessment of language learning in the classroom. Formative self-assessment through the biography and the dossier can be easily integrated into regular everyday classroom activities. Learners can use the checklists on a regular basis to monitor their individual learning outcomes within the progress of the course. Updating samples of work in the dossier can also take place regularly to prove the attainment of new learning goals and to reflect progress in learning. This formative assessment through the biography and dossier should be continuously guided and supported by the teacher and complemented with language tests.

Summative self-assessment through the passport and dossier can also be part of formal assessment. Students can be engaged in systematic summative self-assessment through the dossier. At the end of each course they can select assignments and present them orally in peer groups justifying their choices. Then, they can exchange dossiers with members of the group for peer

assessment on the basis of the 'I can' checklists. The act of being assessed by and assessing peers is one of the deepest learning experiences (Race, Brown and Smith 2005:132) offering learners the opportunity to learn from each other's weaknesses and successes. (For a detailed discussion of the importance of peer assessment to self-assessment see Kostopoulou forthcoming).

Finally, having received peer feedback, they can assess their own dossiers. Feedback from the teacher should also be provided after the peer and self-assessment processes. Then, students can plan their learning in co-operation with the teacher. In the Finnish ELP pilot project for instance, self-assessment for summative purposes took place through the dossier at the end of secondary school 'to explore the potential of the ELP as a school-leaving reporting tool and to give the students an experience of how they might update their ELPs in the future on their own' (Päkkilä 2003:8).

This dialectic scheme of assessment, combining ELP self-assessment with external assessment can feed back into learning and teaching practices, enhance the effectiveness of learner self-assessment and encourage learners to be honest and realistic in their judgements. But ELP self-assessment will have only limited success in enhancing the quality of L2 assessment, learning and teaching unless certain reforms take place so that all pedagogic procedures become compatible with its philosophy.

Necessary reforms and prerequisites

Accommodating ELP self-assessment into formal assessment requires a reconciliation between public examinations and the ELP. This can be achieved by introducing a portfolio-oriented element in final examinations. Learners' ELPs, for instance, could be used for the examination of their oral proficiency (Little 2003a:33). Students could present their portfolios to the examiners and the oral presentation of ELP by individual learners would lend itself to spontaneous questions and answers in the TL (Little 2003b:231). In this case, the ELP could serve as a means of reforming examination systems so that they become consistent with the ELP philosophy and the general educational changes.

Revising examinations to bring them into line with the common reference levels of the CEFR (Figueras, North, Takala, Van Avermaet and Verhelst 2004) is a step towards this direction. Conducting interviews based on group projects between examiners and learners to examine the oral component of languages could be another change to respond to the social-interactive theories of learning. Finally, introducing a double level of assessment, i.e. CEFR level and mark will be a way to recognise students' success and, at the same time, identify their place in the continuum of communicative competence and their level of proficiency in relation to the different possible levels (Goullier 2007:101).

Given that self-assessment in the ELP is conducted against the common reference levels of the CEFR, these levels should be used for the 'constructive

alignment' of L2 curricula (Biggs 2003), that is 'designing of curricula so that the teaching activities, learning activities and assessment tasks are co-ordinated with the learning outcomes' (Kennedy 2007:77). In this way, the common reference levels allow an approach to curriculum, learning, teaching and assessment from the same communicative perspective.

Learner-centred approaches and reflective pedagogical practices are also needed that respond to the complexity of developing teacher autonomy and self-assessment skills in order to help learners develop these skills. Enabling learners to undertake self-assessment successfully requires a greater level of professional awareness and professional sophistication than has usually been the case. This suggests that teacher education should nurture reflective practitioners (Schön 1983) who engage in critical thinking about their teaching, examine the underlying philosophies of their practices, and consider alternative approaches. Reflective journals and teaching portfolios are an effective means of stimulating this kind of reflective enquiry.

A practical example of integrating ELP self-assessment in formal education is offered from English language support classes for immigrant students in Irish post-primary schools.

ELP self-assessment in ESL support in Irish post-primary education

Recognising the need to help students with English as a second language (ESL) access mainstream education and succeed academically, the Irish Department of Education and Science established language support programmes at primary and post-primary levels. According to the official circular (Department of Education and Science 2007), all students are entitled to language support on a withdrawal basis for a maximum period of two years, regardless of legal status. The aim of language support is to promote students' development of English language proficiency so that they can gradually gain access to the Irish post-primary curriculum and benefit from the same educational opportunities as their English-speaking peers, and, in the longer term, secure a place in Irish society. ESL support is, thus, content-based, meaning that L2 learning, teaching and assessment focus on the communicative demands, themes and topics of the Irish post-primary curriculum.

Accordingly, the primary function of language assessment is to support the development of students' subject-specific language skills so that they can become fully integrated into mainstream courses. Pedagogical procedures, including assessment, should also assist students in developing learner autonomy so that they can continue learning independently after the end of ESL support. Continuous independent learning is necessary during schooling for academic success but also throughout students' whole life for their

professional development. Considering the social dimensions of language assessment, the underlying values and ideology of assessment in the context in question should convey messages to learners about what kinds of identity, behaviour and rights are acceptable and valued in the classroom and, by extension, in Irish society.

To serve the pedagogical and social functions of assessment that are appropriate to this context, an assessment framework has been developed for ESL classrooms in which learner self-assessment through the Irish post-primary ELP is an essential component.[1]

Interface between L2 learning, teaching and assessment

In the Irish ELP for post-primary learners (IILT 2004), ESL students' assessment of their language achievement and proficiency is conducted against functional checklists in the language biography which mediate communicative demands, themes and topics of the principal subjects in the Irish post-primary curriculum (e.g. I can understand labels on scientific diagrams and equipment, I can give and explain my views about a story or poem, I can write a clear explanation of a mathematical process etc.). In this way, the school subjects are incorporated fully into the development of language proficiency and learners' access to subject-specific learning is promoted. Thus, the ELP facilitates authentic assessment as it corresponds to learners' needs and it fulfils the function of content-based language assessment in this context.

These subject-specific 'I can do' checklists have multiple functions. They mediate the ESL curriculum to students and suggest teaching objectives necessary for planning ESL provision. Further, a substantial bank of pedagogical materials, tasks and activities has been developed on the basis of the self-assessment checklists to facilitate the achievement of learners' and teachers' objectives. Learners set learning goals, monitor their progress and assess their learning outcomes on the basis of the same checklists that teachers use to define teaching objectives and design pedagogical materials.

What is more, the subject-specific checklists have informed the development of language tests, including rating scales and scoring procedures ('Post-Primary Assessment kit' developed by Integrate Ireland Language and Training and distributed to schools by the DES early in 2009), which can be used in combination with learners' self-assessment results to regularly review their progress in order to identify when they are ready to fully access the mainstream classroom. As a result, the ESL curriculum, learning, teaching, ELP self-assessment and language tests are all oriented to the same functional statements of English language proficiency. The overarching framework of this pedagogical system is the CEFR since it is the point of reference for the ESL curriculum, the pedagogical materials, the ELP self-assessment checklists and the language tests.

Learners' pedagogical growth

The links between assessment and learning are also strengthened in this assessment framework as a result of the consciousness-raising function of self-assessment. By forming judgements about their individual learning process and learning outcomes, learners can 'appreciate their strengths, recognise their weaknesses and orient their learning more effectively' (Council of Europe 2001:192). This renders both the assessment process and the assessment criteria more transparent to learners and helps them to achieve their short and long-term goals more easily. In this respect, self-assessment not only initiates a process of evaluation but also provides a basis for setting goals, identifying strategies and planning further learning.

From this, it follows that self-assessment has the potential to improve learners' strategic control of the learning process, which in turn contributes to the development of strategic competence, understood as a 'general cognitive capacity that we draw on in all kinds of problem-solving behaviour' (Little 1997:15). The process of reflecting on the content and process of learning for purposes of self-assessment is a rewarding intellectual activity in itself, because it cultivates learners' metalinguistic awareness and metacognitive skills which are important for lifelong learning.

The ipsative (rather than normative) nature of ELP assessment, allows learners to evaluate their performance by comparing it with their own previous performance against objective criteria and not with that of other learners in a competitive manner. It thus offers an opportunity for ESL learners to develop their own genuine understanding of what they have learned and how and, as a result, helps them to realise that it is largely their own actions that determine their progress. This awareness of personal control and effectiveness in the learning process fosters self-motivation, which is another significant quality of successful learners (Ushioda 1996) that is promoted through self-assessment in the ELP and is particularly important for sustaining ESL students' long-term learning.

Self-motivation constitutes an intrinsic element of the capacity for language learner autonomy which can be developed through self-assessment in the ELP and it is essential for this group of learners. According to Little (2007), autonomy in L2 pedagogy embraces language learning and language use, two sides of the same coin. He argues that reflection, empowerment and appropriate TL use are the three fundamental principles that need to be operationalised for the development of this capacity (Little 2002:51–52). Self-assessment by definition engages learners in critical reflection and it should be conducted through the medium of the TL. The process of continuous self-assessment has an empowering effect because students develop a personal awareness as learners, which motivates them to take charge of their learning and learn how to direct it more effectively. Therefore, ELP self-assessment

assists ESL students in developing learner autonomy as it creates conditions that incorporate its underlying principles.

Learners' socio-cultural and political growth

The ELP invites a sociocultural view of learners, recognising that they are individuals with a particular identity and members of a particular society and culture. Reflective self-assessment focuses accordingly not only on the content and process of L2 learning but on intercultural competence as well. By documenting their intercultural experiences and cultural awareness in the language biography of the Irish post-primary ELP, ESL learners can recognise that their mother tongue and cultural values and beliefs are respected and valued and that diversity is celebrated. The ensuing feelings of acceptance and sense of belonging are necessary if ESL learners are to feel they are welcome in the school community and form positive self-perceptions.

At the same time, self-assessment in the ELP, conducted through critical reflection in the TL, opens up a window to new cultures and ways of thinking and it provides learners with the opportunity to access a foreign reality and construct a new cultural and social identity which interacts with their old one. This can be understood if one considers that language is identity (encompassing culture, values etc.) and language learning is thus a matter of identity formation. In this respect, reflecting on language and its acquisition in the ELP entails the reconstruction of learners' linguistic identities within the Irish context. The new identity ESL learners gradually build (through L2 acquisition) allows them to relate to Irish social reality and supports their socialisation, which is essential for their social integration into the school community and Irish society.

From a critical sociological perspective, the underlying ideology and values of ELP self-assessment create conditions in the classroom which contribute to a kind of 'political literacy' (Starkey 2002:8) and socially responsible learning. More specifically, the successful implementation of self-assessment can only take place within classrooms where responsibilities are shared by all participants (Little 2003b:228) and learners' contributions are valued. Particularly in the case of ESL learners, ELP assessment responds to the need of 'Critical Language [Assessment]' (Shohamy 2001:373) to 'consider voices of diverse and different groups in multicultural societies'. From this perspective the L2 classroom can be viewed as a citizenship site which functions as a model of democratic society. The redistribution of educational power and the more democratic relationships between teachers and learners necessitated by the practice of self-assessment result in an interface between pedagogic democracy and societal democracy.

In this context, self-assessment is part of what Bernstein (2000:xx) calls, 'pedagogic rights', and learners are given responsibilities and choice which

encourages their active participation. The action-oriented nature of ELP assessment, in its socio-political dimension, gives rise to 'pragmatic' or 'action outcomes', which refer to 'improving people's ability to take initiative and to accept responsibilities on society. They are those capacities that empower the individual to take an active part in and contribute to the community, in the shaping of its affairs and in solving problems' (O'Shea 2003:14). In addition, critical thinking for the needs of self-assessment can contribute to the construction of a 'critical social consciousness' (O'Shea 2003:10) which enables learners as future citizens to examine and analyse power structures and issues that affect their lives and make informed decisions.

The practice of self-assessment in formal education also corresponds to current concerns with education for democratic citizenship as it nurtures learners within a climate of democracy and it equips them with the kinds of knowledge and skills that are necessary for responsible political participation. On these grounds, learners can develop autonomy as socio-political agents, since they are educated as responsible citizens who exercise their political rights within democratic societies and are invited to adopt a critical stance towards public affairs. Nurturing responsible individuals and preparing them to function effectively in society is to the benefit of both the immigrant and the host country population.

From more democratic assessment towards more democratic L2 education

Taking into consideration the general aims of language assessment and testing and some of the current concerns in the field, this paper argued that establishing learner self-assessment through the ELP as an integral part of institutionalised L2 education can enhance the quality of assessment and strengthen its links with learning and teaching. I described how the ELP engages learners in formative and summative self-assessment, explained some of the challenges it entails and discussed issues of assessment quality. I then suggested specific ways of integrating ELP assessment in pedagogical procedures and briefly outlined some of the necessary reforms and prerequisites for successful implementation. I concluded by offering a practical example of an assessment framework from a formal educational context (ESL support in Irish post-primary schools) in which ELP self-assessment has a central role.

Innovations in language assessment such as the one presented here have the potential to raise educational standards, but they can also foster an L2 pedagogy that facilitates learners' self-actualisation through pedagogical, social, cultural and political growth. In this way, L2 education corresponds to the view that Aristotle expressed in *The Politics* that the mission of education is to produce good and virtuous citizens for the polis. This mission

is increasingly important today because nurturing 'good and virtuous citizens' within multicultural societies can help to ensure their harmonious coexistence and instil in them the desire for positive social change.

Acknowledgements

I would like to thank Prof. David Little of the School of Linguistic, Speech and Communication Sciences, Trinity College Dublin for his useful insights and detailed proofreading of the manuscript.

Note

1. The assessment framework described in the present discussion was built by a) Integrate Ireland and Language Training (1999–2008), a government-funded institution of Trinity College responsible for delivering English courses to adult refugees and asylum seekers and supporting ESL provision at the school sector in Ireland, and b) the English Language Support Programme, a research project of the Trinity Immigration Initiative (2007–2010; Trinity College Dublin) which substantially contributed to the provision of ESL support in Irish post-primary schools.

References

Bachman, L and Palmer, A S (1989) The construct validation of self-ratings of communicative language ability, *Language Testing* 6 (1), 14–29.

Bernstein, B (2000) *Pedagogy, symbolic control and identity: Theory, research, critique*, Oxford: Rowman and Littlefield Publishers.

Biggs, J (2003) *Teaching for quality learning at University*, Buckingham: Open University Press.

Blanche, P (1990) Using standardized achievement and oral proficiency tests for self-assessment purposes: The DLIFLC study, *Language Testing* 7 (2), 202–229.

Boud, D (1995) *Enhancing learning through self-assessment*, London and Philadelphia: Kogan Page.

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

Broadfoot, P (2001) Empowerment or performativity? Assessment policy in the late twentieth century, in Phillips, R and Furlong, J (Eds) *Education, reform and the state: Twenty-five years of politics, policy and practice*, London: Routledge Falmer, 136–155.

Brown, S (1990) Assessment: A Changing Practice, in Horton, T (Ed.) *Assessment Debates*, London: Hodder and Stoughton in association with the Open University, 5–11.

Council of Europe (2000) *European Language Portfolio (ELP): Principles and Guidelines,* Strasbourg: Council of Europe.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment,* Cambridge: Cambridge University Press.

Council of Europe (2006) *European Language Portfolio (ELP): Key Reference Documents*, Strasbourg: Council of Europe (DGIV/EDU/LANG, 2006, 4).
Cummins, J (1997) *Biliteracy, empowerment and transformative pedagogy*, viewed 23 August 2008 <www.iteachilearn.com/cummins/biliteratempowerment.html>.
deCharms, R (1968) *Personal causation: The internal affective determinants of behaviour*, New York: Academic Press.
Deci, E L [with Flaste, R] (1996) *Why we do what we do: Understanding self-motivation*, New York; London: Penguin Books.
Department of Education and Science (2007) *Meeting the needs of pupils for whom English is a second language* (Circular 0053), Dublin: DES.
Dickinson, L (1989) Learner Training, in Brookes, A and Grundy, P (Eds) *Individualisation and Autonomy in Language Learning*, London: Modern English Publications, British Council.
Figueras, N, North, B, Takala, S, Van Avermaet, P and Verhelst, N (2004) *Manual for relating language examinations to the 'Common European Framework of Reference for Languages' (Preliminary pilot version)*, Strasbourg: Council of Europe.
Goullier, F (2007) *Council of Europe tools for language teaching: Common European Framework and Portfolios*, Strasbourg: Council of Europe.
Holec, H (1985) Self-assessment, in Mason, R (Ed.) *Self-Directed Learning and Self-Access in Australia: From Practice to Theory*, Melbourne: Council for Adult Education.
Integrate Ireland and Language Training (IILT) (2004) European Language Portfolio for post-primary learners, Dublin: IILT.
Kennedy D (2007) *Writing and using learning outcomes*, University College Cork, Ireland.
Kohonen, V (2002) The European Language Portfolio: From portfolio assessment to portfolio-oriented language learning, *Developments in reflective language learning and self-assessment, Section 2*, in Kohonen, V and Kaikkonen P (Eds) *Quo vandis foreign language education?* Tampere: Tampere yliopiston julkaisusarja, 77–95.
Kostopoulou, S (forthcoming) *Learner self-assessment of foreign language proficiency and the European Language Portfolio: Towards a new assessment culture*, Centre for Language and Communication Studies Occasional Paper, Dublin: Trinity College Dublin.
LeBlanc, R and Painchaud, G (1985) Self-assessment as a second-language placement instrument, *TESOL Quarterly* 19 (4), 673–686.
Little, D (1997) Strategic competence considered in relation to strategic control of the language learning process, in Holec, H, Little, D and Richterich, R *Strategies in language learning and use*, Strasbourg: Council of Europe, 9–37.
Little, D (2002) We're all in it together: exploring the interdependence of teacher and learner autonomy, in Karlsson, L, Kjisik, F and Nordlund, J (Eds) *All Together Now* (Papers from the 7th Nordic Conference and Workshop on Autonomous Language Learning, Helsinki, September 2000), Helsinki: University of Helsinki, Language Center, 45–56.
Little, D (2003a) *Languages in the post-primary curriculum: a discussion paper* (Discussion paper), Dublin: National Council for Curriculum and Assessment of Ireland.
Little, D (2003b) Learner autonomy and public examinations, in Little, D, Ridley, J and Ushioda, E (Eds), *Learner autonomy in the foreign language*

classroom: Teacher, learner, curriculum and assessment, Dublin: Authentik, 223–233.
Little, D (2005) The Common European Framework of Reference for Languages and the European Language Portfolio: Involving learners and their judgements in the assessment process, *Language Testing* 22, 321–336.
Little, D (2007) Language learner autonomy: Some fundamental considerations revisited, *Innovation in Language Learning and Teaching*, 1 (1), 14–29.
Little, D and Perclovà, R (2001) *European Language Portfolio: A guide for teachers and teacher trainers*, Strasbourg: Council of Europe.
McNamara, T (2001) Language assessment as social practice: challenges for research, *Language Testing* 18 (4), 333–349.
McNamara, T (2008) *Recognizing the Other: Language assessment, immigration and citizenship*, presentation at the ALTE 3rd International Conference, 10–12 April, Cambridge.
McNamara, T and Roever, C (2006) *Language testing: the social dimension*, Oxford: Blackwell.
McNamara, T and Shohamy, E (2008) Language tests and human rights, *International Journal of Applied Linguistics* 18 (1), 90–95.
O' Shea, K (2003) *A glossary of terms for education for democratic citizenship: Developing a shared understanding*, Strasbourg: Council of Europe, viewed 23 August 2008 <http://www.coe.int/t/dg4/youth/Source/Resources/Documents/Glossaire_ECD_en.pdf>.
Oscarson, M (1997) Self-assessment of foreign and second language proficiency, in the *Encyclopedia of Language and Education* 7 (Language Testing and Assessment), Dordrecht: Kluwer Academic Publishers, 175–187.
Oskarsson, M (1978) *Approaches to self-assessment in foreign language learning,* Strasbourg: Council of Europe.
Oskarsson, M (1984) *Self-assessment of foreign language skills: a survey of research and development work,* Strasbourg: Council of Europe.
Oskarsson, M (1989) Self-assessment of language proficiency: Rationale and applications, *Language Testing* 6, 1–13.
Päkkilä, T (2003) The Finnish ELP pilot project for upper secondary schools, in Little, D (Ed.) *The European Language Portfolio in use: nine case studies*, Strasbourg: Council of Europe, 7–18.
Race, P, Brown, S and Smith, B (Eds) (2005) *500 Tips on Assessment* (second edition), London: RoutledgeFalmer.
Ross, S (1998) Self-assessment in second language testing: A meta-analysis and analysis of experiential factors, *Language Testing* 15, 1–20.
Satterly, D (1989) *Assessment in schools* (second edition), Oxford: Basil Blackwell.
Schön, D (1983) *The reflective practitioner*, London: Basic Books.
Shohamy, E (1998) Critical language testing and beyond, *Studies in Educational Evaluation* 24 (4), 331–345.
Shohamy, E (2001) Democratic assessment as an alternative, *Language Testing* 18 (4), 373–391.
Starkey, H (2002) *Democratic citizenship, languages, diversity and human rights*, Strasbourg: Council of Europe.
Ushioda, E (1996) *The role of motivation*, Dublin: Authentik.

Section Three
Reflections on the impact of testing on stakeholder constituencies

16 Standards-based assessment in the US: social and educational impact

Micheline Chalhoub-Deville
University of North Carolina at Greensboro

Abstract

The paper addresses the social and educational impact of an educational reform movement in the US, called No Child Left Behind (NCLB) (2001). Within NCLB, the paper focuses on Title III standards-referenced assessments (SRAs), which are intended to measure the progress and attainment of English language learners (ELLs) in terms of *academic* English language proficiency. In discussing educational and social impact, the paper addresses three primary issues. First, the paper elaborates how the L2 construct is operationalised in terms of NCLB Title III SRAs and the extent to which these assessments yield scores that can be interpreted and used to help ELLs. Second, the paper discusses emerging views on the responsibility of test developers and test users with regard to the documentation of impact. Given the increasingly confounded role of test developers and users, the paper calls for explicit and upfront negotiation of roles, expectations, and activities with regard to impact research. Third, the paper calls for the expansion of the traditional conceptualisation of impact research and introduces Social Impact Analysis (SIA) to emphasise the need for anticipatory impact research to inform and guide policy formulation.

Introduction

The paper addresses the social and educational impact of an educational reform movement in the US, called No Child Left Behind (NCLB) (2001). In addressing impact issues, the paper focuses primarily on a particular segment of NCLB standards-referenced assessments (SRAs), i.e. those addressed in Title III of the Act. Title III SRAs are intended to measure the progress and

attainment of English language learners (ELLs) in terms of *academic* English language proficiency.

No Child Left Behind ushers in significant changes to testing in the schools in the US such as the testing of all students, and the introduction of sanctions to provide corrective actions in schools that perform inadequately by not meeting annual targets over a period of time. In general, NCLB is viewed in a negative light. The following is an excerpt from a letter written by a group of educators, parents, and citizens called the Community Dialogue on Education (2008). The letter is written to legislators in North Carolina, the state where I live and work.

> Although we believe that some of the goals of that legislation were laudable, we were convinced that features of the legislation would not only cause NCLB to fail to meet its goals, they will actually cause a deterioration of the Country's public schools. In fact, these deficiencies were so dramatic that we fully understand why some critics of NCLB said that the goal of NCLB was to shut down public education.

This is quite a dramatic statement but represents the sentiment of many educators and lay people in the US who know the financial and educational costs of the legislation (see Hulbert 2007, Menken, 2006, 2008).

Despite NCLB's above mentioned shortcomings, a positive outcome of the Act has been its attention to the ELL student population, which used to be exempted from statewide achievement tests and whose needs were subsequently not met in the schools. ELLs who were excluded from a state's testing requirements were marginalised in the schools (Bailey and Butler 2003, Chalhoub-Deville and Deville 2008). NCLB, however, calls for formal attention to the instructional and assessment needs of ELLs. By including ELLs in high-stakes tests, educators and schools are now compelled to improve this group's learning opportunities and increase their educational achievement.

ELLs represent an important segment of students in US schools. According to Francis, Rivera, Lesaux, Kieffer and Rivera (2006), US schools include about 10 million ELL students. Roughly half of these students are identified as limited English proficient (LEP), which is the federal government's designation for this student population. Schools receive federal funds to help attend to the needs of ELLs, a rapidly growing population within schools. The ELL group is on the rise. In the period between 1979 and 2003, schools witnessed a 169% increase in the ELL population and by 2015, it is expected that this group will represent 30% of school-age students. ELLs represent a diverse group. In terms of first language alone, these students speak over 400 languages. The majority of them, about 70%, however, speak Spanish as their first language.

Overview of impact

When discussing the influence of testing and testing practices on individuals, groups, and society, we are referencing what is known in the field as 'impact', 'washback', 'backwash', or 'consequences'. These terms are used more or less interchangeably. Given the widespread use of the term impact, however, and for ease of communication, I will predominantly use this term throughout the paper. Language testers have embraced the term impact. The field includes substantial publications in this area (Alderson and Wall 1993, Chalhoub-Deville and Deville 2006, Cheng 2005, 2008, Garcia 2003, Green 2007, Hawkey 2006, Hamp-Lyons 1997, Shohamy 1996, 2001, Spolsky 1997, Wall 1996, 2005, and Wall and Alderson 1993).

In the measurement field and prior to Messick's seminal paper on validity in 1989, discussion of impact was implicit. It was Messick who made the discussion of 'impact' and 'consequences' explicit and integral to validation. To a large extent, measurement professionals are divided in terms of their support for consequences as integral to validation and hence the responsibility of test developers. This argument is beyond the scope of this paper. Those interested in reading more about this topic are referred to Chalhoub-Deville (2009).

For the purposes of the present paper, it is sufficient to focus on the *Standards for Educational and Psychological Testing* (1999), coauthored by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The *Standards* (AERA, APA, NCME 1999) could be said to represent the official position of the measurement profession. With regard to impact and validation, the *Standards* (1999:16) state:

> Evidence about consequences can inform validity decisions. Here, however, it is important to distinguish between evidence that is directly relevant to validity and evidence that may inform decisions about social policy but falls outside the realm of validity. . . Thus evidence about consequences may be directly relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components. Evidence about consequences that cannot be so traced. . . is crucial in informing policy decision but falls outside the technical purview of validity.

Essentially, the *Standards* (AERA, APA, NCME 1999) acknowledge that consequences/impact that could be traced to aspects of the construct are directly related to validity and demand attention. The *Standards* identify two construct-related threats to validity: construct underrepresentation and construct irrelevant variance. An example of construct underrepresentation is a test of communicative language proficiency or communicative language use test that

does not include a measure for speaking. Construct irrelevant variance pertains to a situation such as a computerised writing task where students do not have the necessary typing skills. A case could be made that the ability to type is irrelevant to the writing construct and that compromises the quality of the information resulting from the test scores. In short, impact investigations are warranted in terms of matters pertaining to the construct, i.e. obtaining valid scores that reflect performance with respect to the construct of interest.

Another important aspect of the impact discussion is the delineation of roles and responsibilities. The *Standards* (AERA, APA, NCME 1999:112) state that 'the ultimate responsibility for appropriate test use and interpretation lies predominantly with the test user'. While the *Standards* acknowledge that test developers are responsible for threats to the construct, they assert that test users are responsible for test score use. This is a very complex issue and the discussion, later in the paper, shows that matters have been rendered even more complex with NCLB testing practices. Briefly, complications arise because of the increasingly less differentiated roles of test developers and users under NCLB.

Yet another important consideration with regard to impact in language testing and the measurement field at large is the reactive nature of research investigations. Absent from the second language (L2) literature are attempts to study or push for impact to inform policy making before a large-scale and high-stakes assessment is implemented. A useful concept to consider in this regard is Social Impact Analysis (SIA), which is popular in different areas of study such as anthropology and environmental science. SIA emphasizes the importance of research into impact before a policy and/or a test is in place. SIA is a proactive rather than a reactive research approach to the examination of impact and consequences.

In the present discussion of educational and social impact, the paper addresses three primary issues. First, the bulk of the paper elaborates how the L2 construct is operationalised in terms of NCLB Title III assessments and the extent to which these assessments yield scores that can be interpreted and used to help ELLs. After all, educational reform is intended to help ELLs attain higher levels of proficiency in English and other content domains. Second, the paper discusses emerging views on the responsibility of test developers and test users with regard to the documentation of impact. Given the increasingly confounded role of test developers and users, researchers are calling for explicit and upfront negotiation of roles, expectations, and activities with regard to impact research. Third, the paper introduces SIA to emphasise the need for anticipatory impact research to improve policy making and educational reform objectives. The purpose in this section is to point out the dearth of impact investigations beyond the dominant reactive policy research that document the influence of policies only after the fact. The section calls for anticipatory research that informs and guides policy formulation.

Impact: Academic English language proficiency

President George W Bush continued the educational reform movement that was advanced by his father, President George H Bush who introduced Goals 2000, also advocated by President Bill Clinton, who tried to institute national standards and national assessments (for a detailed discussion of this point, see Chalhoub-Deville and Deville 2008). President George W Bush called for state-based standards and SRAs, which are aligned with the standards. In terms of ELLs, NCLB Title III requires states to use SRAs that document the progress and attainment of ELLs in terms of *academic* English language proficiency.

Published literature. Many would agree that the notion of differentiated language use goes back to Cummins (1981) who elaborated differences between Basic Interpersonal Communicative Skills (BICS) and Cognitive/Academic Language Proficiency (CALP). BICS refers to language used in everyday, informal situations whereas CALP refers to language employed in the schools, in the classroom and other formal educational settings. CALP refers to the language observed in textbooks and examinations. It also refers to the language associated with particular content areas such as maths, science and social studies.

BICS and CALP are utilised in different contexts and thus have features particular to each of them. According to Cummins (2001:123–124), 'academic language entails vocabulary that is much less frequent than that typically found in interpersonal conversation, grammatical constructions that are unique to text and considerably more complex than those found in conversation, and significant cognitive processing demands that derive from the fact that meanings expressed paralinguistically in conversation (e.g. gestures, facial expressions, intonation, etc.) must be expressed linguistically in text'. This delineation, while helpful in the distinctions it draws between everyday and academic language use situations, functions, and other features, provides insufficient information to inform the development of test specifications.

A review of the literature published before 2001, when NCLB was signed into law, shows that researchers had not yet provided any detailed operationalisation of language observed in textbooks, classrooms, subject areas, schools, and tests at different school levels (see Butler and Stevens 2001 and Bailey and Butler 2003). As expected, this lack of detailed information regarding CALP presented test developers with a great challenge. It is quite daunting for test developers to construct a complex battery of assessments such as that required by NCLB (e.g. all four modalities at all grade levels) given an unreasonably short timeline. The difficulties were exacerbated, however, when the CALP construct had yet to be adequately investigated and defined.

Standards. Given the dearth of information regarding the elaboration of ALP in the published literature, test developers looked to language and content standards to help them define the ALP domain. Multiple compendia of standards were available (e.g. professional, national, and statewide language, maths, science, and English language arts). The most obvious set of standards to consult for ALP is the *ESL Standards for Pre-K-12 Students* (TESOL 1997). These *ESL standards* differentiate between social language use, i.e. Goal 1 and academic language use, i.e. Goal 2. However, these standards proved to be an unreliable foundation on which to build tests. Bailey and Butler (2003) note that considerable overlap exists in terms of language and task expectations for these two goals. This overlap in social and academic language use and tasks renders the development of specifications for measuring academic language quite challenging. Similar to the published literature, the L2 standards were not useful in terms of informing and guiding the development of the academic language proficiency assessments. (It is important to point out that since 1997 new standards have become available. The *ESL Standards* (TESOL 2006) focus entirely on elaborating the language proficiency construct in the content domains of science, maths, English language arts, and social studies at different grade levels. However, these standards were published well after states began administering ALP tests.)

The content standards from the academic subject disciplines (e.g. maths, science, etc.) proved to be even less helpful for formulating ALP specifications. Extracting the language features embedded in the content standards presented a significant challenge. It is reasonable to state that the content standards were developed with an eye to communicate the knowledge and skills that students are expected to know in the content domains of science, maths, and social studies, and not to delineate the language features ELL students need to know in order to perform in these content areas. In short, language test developers had access to content standards but they could not simply take these standards and develop ALP specifications. An important but missing step in the process is the need to analyse and translate the content standards into language features that ELLs need to communicate in that subject area. Moreover, given that ALP pertains to language in different subject area domains, research is needed to extract and combine the language and task features mandated in the different subject area standards and to align these with features observed in language standards. Researchers have started work in this area. For example, Bailey, Butler and Sato (2005) have been successful in developing systematic, rigorous standards-to-standards linkages that involve both language and content standards. However, what is critical to the present discussion is the absence of such information and research at the time that test developers were embarking on the development of Title III SRA specifications.

English language arts (ELA). One clear misrepresentation of ALP, which can lead to negative consequences, is the confusion of the ALP construct with ELA. According to Kieffer, Lesaux and Snow (2007), 'in most states, the expectations for language minority students are derived from English Language Arts'. ELA standards describe what students should know and be able to do in their English language and literacy school work. While it is quite appropriate to consider ELA when developing Title III ALP SRAs, basing the entire assessment of ELLs on ELA standards constitutes a flagrant construct underrepresentation. ELLs need to learn language as used in maths, science, social studies, and other areas as well. A focus on ELA deprives ELLs of the opportunities they need to develop knowledge, skills, and abilities to become proficient academic language users who are able to perform in different content domains.

Test results. Educational reform has a greater chance of success if it involves key stakeholders such as educators and parents of ELL students. Yet test results are too often reported in formats designed by measurement professionals without adequate consideration of other audiences. It is important to simplify and present test results in ways that non-measurement experts find useful. Test developers should consider relying more on graphs and tables to try to communicate information regarding test scores. It is equally important to report information in multiple languages so parents can stay abreast of their children's educational progress. Test developers of the ACCESS for ELLs by World-Class Instructional Design & Assessment (WIDA), a consortium of 15 states with the Center for Applied Linguistics, have embraced this approach and they communicate test score information in multiple languages.

Another consideration to help enhance test score interpretation and use pertains to reporting disaggregated scores for ELL students. Currently, NCLB requires states to report scores separately for ELL students in the content area achievement. While this requirement should be seen favourably because it focuses attention on the performance of this specific group, it falls short in terms of capturing necessary information to monitor growth for these students. Viewing ELL students as a homogenous group masks important differences with regard to ELLs' language development and yields information of limited instructional use. The second language acquisition literature identifies several key variables that impact significantly the development of language learners. Variables such as the comparability of ELLs' first language (L1) to English, ELL students' L1 and L2 literacy, the number of years ELLs have spent in their homeland and in US schools, among others, are critical considerations to better understand ELLs' progress in English and to provide them with relevant instruction (see Freeman and Freeman 2001, Fradd, McGee and Wilen 1994, Herrera and Murry 2005). California is a good example of a state that has moved beyond documenting only

aggregated scores for ELLs. They also document performance according to many of the above mentioned background variables. Monitoring performance according to key background variables that impact ELLs' language development should become standard practice for all, if we are serious in our efforts to use test scores to help students in the classroom.

Once ELLs attain a specified level of identified language proficiency, they are reclassified and moved out of the designated ELL group in terms of reporting for NCLB. Many question this practice. For example, according to Francis et al (2006:5), 'many ELLs who are no longer formally designated (ELL, LEP) continue to struggle with academic text and language'. With proper language support, ELLs may reach a point where they have attained an adequate level of ALP to be able to function comfortably in the school environment and where language is no longer a barrier to be meaningfully engaged in the curriculum. However, as ELLs move into more advanced grade levels and the academic language gets more complex, they may, once again, fall behind in terms of their ALP in English. It is, therefore, important for educators to continue to monitor the performance of ELLs beyond reclassification to ensure that their language skills do not become a hindrance to their interaction with and access to academic content.

Individual educational plans. Research into classroom practices shows that NCLB has impacted classroom activities considerably. According to a year-long classroom observation and teacher interview study of 10 New York high schools, Menken (2006) reports that even in bilingual classes teachers were emphasising instruction in English at the expense of other languages. Additionally, where first language was being used, it was used primarily to help prepare students for NCLB tests. Menken also observed, similar to what was reported earlier, that teachers seem to be targeting ELA at the expense of the ESL curriculum.

Menken (2006:538) remarks that 'these issues have arisen because NCLB is essentially a "one-size-fits-all" educational reform into which ELLs are now awkwardly being included'. Menken and other researchers (e.g. Chalhoub-Deville 2008a, 2008b, Menken 2008, Rivera 2008) call for individually tailored instructional and assessment programmes that better address the needs of each ELL student. Such individualised programmes have been available for students with special educational needs. Individual Educational Programmes (IEPs) specify the types of instruction and accommodations a student with special needs requires in the classroom and on tests. IEP-like practices for ELLs would call attention to the individual needs of this population of students according to their background variables, e.g. L1 and L2 literacy, L2 proficiency and literacy level, age, grade, language of instruction, etc. (I strongly emphasise that my intention here is not to lump ELLs with special education students. I am fully aware of the prevalent malpractice of identifying ELLs as students with learning and behavioural disabilities,

ignoring how L2 proficiency impacts the achievement and academic engagement of ELL students. My intent in pointing to students with special needs is to highlight accepted educational practice that could be adopted to fit ELLs' learning and testing needs.)

Summary. This section of the paper addresses the impact and consequences of the NCLB educational reform movement in terms of validity, i.e. the interpretation and use of the L2 construct. The arguments presented document the lack of adequate representation of the ALP construct in the published literature at the time that NCLB called for the development of Title III SRAs. Typically, the L2 construct was discussed and elaborated exclusively in terms of social language use. The dearth of information about academic language use forced states and test developers to undertake the elaboration of the ALP construct on their own in order to formulate test specifications. Additionally, the section discusses serious challenges when elaborating the ALP construct in terms of language and content standards. Some have opted to use English ELA as proxy for ALP. The paper argues that the operationalising of the construct in terms of ELA results primarily in construct underrepresentation (ignores important ALP features such as maths and science language). The section calls for explicit links among the various language and content standards in order to more systematically align SRA specifications with standards.

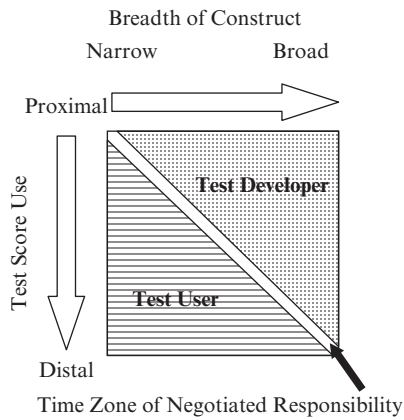
The section draws attention to the need to improve the quality of information gleaned from test results. Any improvement in the operationalisation of ALP in Title III SRAs is likely to be compromised if unaccompanied by efforts to improve communication of test results. We need to utilise more targeted visuals and multiple languages to accommodate the needs of different stakeholder groups, especially educators and parents. For educational reform activities to be meaningful, test scores need to convey clear and understandable information to multiple audiences about the meaning of test scores and what they denote in terms of instructional practices.

Concern with appropriate score interpretation and use demands for the monitoring of test results not just for ELLs as a monolithic group but also for ELL subgroups in order to better accommodate the learning needs of these students. Current NCLB requirements ignore important differentiating background variables that impact ELLs' growth and achievement. Such information is germane to action that would enhance the learning opportunities for these students. More strongly, the arguments presented in this section call for introducing IEP-like practices for ELLs. IEPs will build on research into the role that background variables play with regard to the ELLs' ALP development. Finally, the section calls attention to the need to monitor ELLs beyond reclassification to ensure that weak language skills do not re-emerge as a hindrance for these students to be meaningfully engaged in academic materials.

Impact: responsibilities of test developers and users

The second part of the paper discusses educational and social impact in terms of the responsibilities of test developers and test users. As indicated earlier, The *Standards* (AERA, APA, NCME 1999) hold test users to be ultimately responsible for valid test score interpretation and use. This statement is in some ways unrealistic in its unqualified assertion, because it relies on a clear distinction between test developers and users. The present discussion relies heavily on a framework (see Figure 1) recently advanced by Nichols and Williams (2008). The framework, introduced as part of a presentation at the 2008 annual meeting of NCME, includes three variables, the *Breadth of Construct, Test Score Use,* and *Time* that shape responsibility considerations for test developers and users with regard to validation and impact. The framework is innovative in its fluid depiction of test developers' and users' responsibilities with respect to the dynamics of the three variables. The fluidity of the framework is quite useful in the NCLB testing context, where it is difficult to neatly differentiate test developers from test users. Under NCLB, government agencies, typically conceived of as test users, also play the role of test developers. Government agencies are directly involved in, if not prescribing, test specifications and it is they who shape score interpretation and use.

Figure 1 Test developers' and users' responsibilities



Construct. One key consideration of responsibilities is the *Breadth of Construct* under investigation, i.e. the extent to which the construct is narrowly or broadly defined. It is argued that the broader the elaboration of the construct, the more extensive test developers' responsibility is in terms of providing evidence to support the expansive interpretation and use of the

related scores. The position represented in this variable is widely accepted and endorsed as professional practice for test developers by the *Standards* (AERA, APA, NCME 1999).

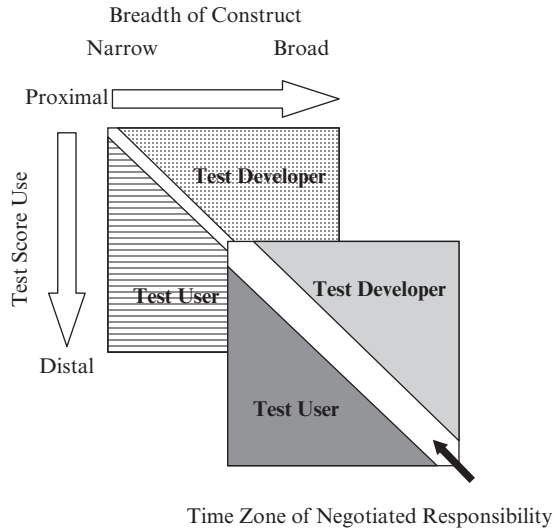
Use. Another consideration is *Test Score Use,* which denotes the extent to which a given testing instrument is utilised as originally planned by its test developers. As depicted in Figure 1, this is not a dichotomous variable. Use is represented along a continuum, which is anchored by *Proximal* and *Distal*. The farther away from test developers' explicitly stated test purpose and use, i.e. the proximal end of the continuum, the more test users are held responsible for undertaking research to document the appropriateness of score interpretation and use. In other words, the extent to which test scores diverge from test developers' intended purposes is directly related to the extent to which responsibility shifts from test developers to test users.

Time. Similarly, time elapsed since a given test has become operational with a given purpose, results in a shift of responsibility. Nichols and Williams (2008) contend that as time passes, test developers must be held accountable for instituting validation research focusing on the unintended interpretations and uses of the test. The argument basically states that despite the fact that users are utilising a given test for unintended purposes over time, test developers cannot continue to regard those purposes as unintended. What was once unintended and new, with the passage of time, is rendered common practice. At that point, it becomes important for test developers and users to enter the so-called Zone of Negotiated Responsibility (ZNR) to reach an understanding of who is responsible for what in terms of validation.

Visually, the need to reconsider responsibility for unintended test score interpretations after a period of time has elapsed and such practices have become commonplace is depicted in Figure 2, as an expanded ZNR area. Whereas Figure 1 shows a narrow ZNR, indicating less of a need to negotiate responsibility for research to support unintended purposes (the responsibility lies with the users), Figure 2 depicts an expanded ZNR to signify the need for test developers and relevant test users to revisit and negotiate those responsibilities.

Multiple factors support the argument for the reengagement of test developers. For example, test developers may be benefiting financially from the expanded use of the test. Additionally, responsible professional practices necessitate that test developers take action to bring to an end such unintended practice or convincingly confer with users on how best to support new and test purposes. In summary, with the passage of time, developers need to re-engage in validation and impact research for practices, which were not originally intended. The passage of time necessitates revisiting responsibilities and negotiating research resources and tasks between test developers and relevant users.

Figure 2 Test developers' and users' responsibilities over time



Roles and responsibilities. As indicated earlier, the *Standards* (AERA, APA, NCME 1999) regard test users as ultimately responsible for test score interpretation and use. The roles of developers and users and concomitantly their respective responsibilities are, however, conflated under NCLB practices. For example, traditionally, schools, districts, federal and state agencies have played the role of test users. Under NCLB, however, these groups are also acting as test developers. Nichols and Williams (2008) elaborate this point as follows: 'the federal government has passed federal legislation specifying features of the test, the state legislature has passed state legislation specifying additional features of the test and the state board of education may have approved the test design'. These agencies are not simply making use of information from test results but are dictating the design of the test. Accordingly, agencies, which are increasingly assuming roles typically reserved to test developers, need to acknowledge their shared responsibility in terms of a research agenda focused on investigating the impact of NCLB tests, including a focus on the construct as well as the educational and social impact.

NCLB practices necessitate that all groups involved in test design enter the ZNR to negotiate their responsibilities for validation and impact research. It is not clear, however, that government agencies are likely to have the resources and professional expertise needed to engage in such research. It is incongruous, however, for agencies to dictate test design without taking interest in and responsibility for the needed research. Resources need to be

made available not only for test development but also for the related validation and impact research. Additionally, these responsibilities need to be addressed and negotiated upfront. In summary, unqualified assertions that a group be held responsible for validation and impact research is not a useful position. Practical frameworks involving co-operative research are needed to delineate circumstances and relationships that outline workable arrangements for conducting necessary validation work.

Impact: proactive research

The last part of this paper deals with a concept addressed in fields such as anthropology, sociology, tourism, and environmental science, i.e. Social Impact Assessment (SIA). SIA is intended 'to help individuals, communities, as well as government and private sector organizations understand and be able to anticipate the possible social consequences on human populations and communities of proposed project development or policy changes' (Burdge 2007:2). Unlike most, if not all, L2 impact research, which tends to be reactive in nature to policy implementation, SIA emphasises the need to engage in impact investigation *a priori*. SIA calls for research before a policy is formulated and put in place. Proponents of SIA caution against designing policies without seriously taking into account potential, intended and unintended, implications. The practices of SIA entail a systematic, proactive research agenda to understand possible educational, linguistic, cultural, and social impact of proposed policy components in order to improve policy making and implementation.

SIA facets. SIA initiatives include core features, which are highlighted by the Interorganizational Committee on Guidelines and Principles for Social Impact Assessment (see Barrow 2000, Burdge, Fricke, Finsterbursch, Freudenburg, Gramling, Holden, Llewellyn, Petterson, Thompson and Wiliams 1995). SIA stipulates designing a systematic plan to investigate the impact of proposed changes. The plan should address relevant stakeholders (individuals, groups, and organisations), especially those likely to be affected and/or at risk. 'If the SIA process is effective, dispassionate and thorough, it should identify and help to counter attempts to manipulate development to serve special interest groups' (Barrow 2000:37). SIA emphasises the evaluative part of intended and unintended change. It posits questions such as: What dis/advantages are garnered and by whom? How are different groups likely to react to different plans of action? Is the plan still workable despite the challenges?

The SIA plan attends to the need to anticipate negative impact and suggest mitigation plans. In other words, SIA emphasises the need to move beyond impact identification. It calls for recommendations regarding 'actions, measures and ongoing monitoring' (Barrow 2000:36) to avoid or moderate the

negative impact likely to take place if policy implementation moves forward. SIA highlights the importance of communicating impact considerations at the policy planning stage. 'Provide feedback on social impacts to project planners and identify problems that can be solved with changes to the proposed action or alternatives' (Barrow 2000:9). Finally, SIA underscores the importance of engaging policy makers with concrete impact findings regarding proposed changes.

With respect to the last point of engaging policy makers, it is important to point out that it is not sufficient to simply inform policy makers about our research findings. Instead we need to work with policy makers. Typically, we researchers confine ourselves to the written document of our work. We do not necessarily see it as inherent to our professional responsibilities to lobby and/or engage legislators and policy makers. Our work is likely to fall short in terms of its intended power to inform and change, if it does not move beyond the publishing/presenting phase. An alternative, as well as a more responsible professional position, is to design plans as part of our research agendas that involve working with policy makers to inform more sound policy making and testing practices. It is understandable that many professionals are constrained (e.g. because of promotion and tenure requirements, job descriptions, etc.) in terms of activities in which they can engage. Consequently, it is critical, if not as individuals then as professional organisations, to be more actively engaged in collaborations to inform policy makers about proposed policies. Organisations such as the International Language Testing Association, as well as regional and national associations, need to become more policy-engaged. Currently, the few organisations engaged in helping shape policy making are the exceptions to the norm.

SIA and NCLB. Given the above description of SIA, it is pertinent to ask: What are some examples of adverse practices that could have been avoided had SIA been part of the policy formulation process? To address this question the following examples are offered. First, within the NCLB, once ELL students are designated proficient, based on their test scores on Title III SRAs, they are moved out of the ELL category, meaning that their test scores are combined and reported with those of the language majority students. One of the stated goals of the Title III legislation is to document the progress of all ELLs toward ALP. The NCLB framework, however, called for ELL students who attained a designated degree of proficiency to be reclassified. They were no longer considered ELLs, no longer required to take the Title III language assessment, and, consequently, their progress was no longer documented. In other words, the better performing students were constantly being reclassified leaving only the lower achieving students in the ELL group, whose scores were then reported to the federal government to document the progress of the ELL population in terms of attaining English language proficiency. Schools and states complained to the federal government that the reporting system

rendered it impossible for ELLs to demonstrate progress when the composition of the group was constantly changing and the proficient students were removed. In response to these complaints, the US Department of Education changed the requirements to allow states to report on the performance of proficient ELL students for up to two years after they have been reclassified. This is a good example of a practice that could have been predicted and easily avoided before had SIA been part of the policy formulation process.

Another example that illustrates how negative impact could have been easily mitigated is the mandating of Title III SRAs when language and content standards were not yet available/usable (see Chalhoub-Deville and Deville (2008) for a discussion on the state of ELL standards in the US). As part of SIA research, a survey of available language and content standards could have been undertaken. Subsequently, findings would have documented the absence and/or impoverished state of such standards to be useful for developing test specifications. NCLB policies informed by SIA research would have allotted states more lead time to develop appropriate standards before test developers were asked to develop operational Title III ALP tests.

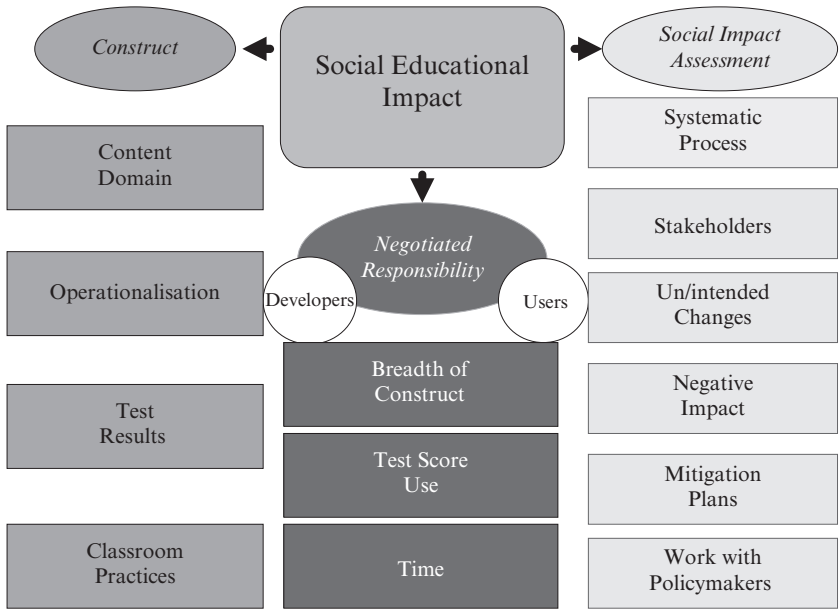
Yet another more complex example is the lack of foresight with regard to how the decentralised approach to education in the US would impact the establishment of performance standards for ELLs. Under NCLB states are free to decide on what students are expected to know, develop their own standards, choose or construct their statewide assessment procedures, and set their individual annual proficiency targets. The result is that each state defines who ELLs are, how they are tested, and what constitutes their being classified as 'proficient'. This decentralised approach has contributed to what some have called educational confusion (Hulbert 2007:12). Hulbert writes in the *New York Times Magazine:* 'But the test mess could be . . . a chance to consider the case for national standards and a single national exam. . .'. Such action, however, is considered a serious threat to the US longstanding and widely accepted tradition of state and local district control over education. Systematic research into policy impact could have predicted the current chaos. Responsible policy making that involves SIA could have suggested alternatives to help mitigate this chaotic situation.

In conclusion, educational reform cannot be based on armchair policies. Educational reform should include systematic plans to investigate the potential impact of policy features. It should anticipate and address intended and unintended impact for various stakeholders. Additionally, the repeated shortsightedness of educational reform policies such as NCLB should serve as a call for researchers to engage in proactive impact research, i.e. SIA, to help shape more sound educational reform policies. While the push to incorporate SIA into our research and policy formation is no guarantee for better educational reform outcomes, it is likely to improve what seems to be an undisciplined process.

Summary and conclusion

I was asked to address in this paper the social and educational impact of educational reform in US schools. In response, the paper focuses on impact issues as they relate to NCLB Title III SRAs, which are designed to measure ELLs' progress and attainment of ALP in English. The paper elaborates on educational and social impact of Title III SRAs primarily with regard to three topics: the ALP construct and validity, roles and responsibilities of test developers and users, and SIA. Figure 3 provides an outline of the issues addressed for these three topics.

Figure 3 Impact research considerations



First, in terms of the construct, the paper reports on the dearth of systematic elaborations of the ALP construct in the L2 literature at the time that these SRAs were being developed. Test developers undertook their own research to operationalise ALP and created their own systems for aligning tests with language and content standards, as mandated by NCLB. These circumstances resulted in diverse representations of the construct across different Title III SRAs. Some developers, for example, opted to rely on ELA standards to operationalise ALP, which constitutes a threat to the ALP construct and compromises the interpretation and use of SRA scores. Another challenge that NCLB Title III SRA developers faced is the standards-to-

standards linkages needed to create a common academic English language domain. Systematic procedures to establish these linkages did not exist and they remain to this day in their infancy. Given the continued popularity of standards-based instruction and assessment, it is critical that researchers pay more attention to developing standards-to-standards linkages procedures.

Additionally, discussion in the paper raises concerns regarding the reporting of test results to major stakeholders in educational reform, e.g. teachers and parents. The argument calls for incorporating more visuals and translating results into languages parents can understand to improve communication with groups whose role is critical for policies and testing practices to have a chance to succeed. The section emphasises the need to be mindful of the impact of NCLB on classroom practices, reclassification of students, and the homogenisation of ELL groups. The concerns with regard to each of these variables are elaborated. Finally, an argument is made in favour of IEPs, i.e. individual educational programmes that tailor to and accommodate the diverse learning and assessment needs of ELL students. In conclusion, the section makes the case that if the NCLB educational reform is to be taken seriously in terms of its attempt to improve instructional practices for ELL students, several factors need to be attended to in terms of best practice. Those addressed in the section represent key variables that cannot be ignored.

Documentation of impact is also discussed in terms of the responsibilities of test developers and test users. Concerns are raised that static representations of these responsibilities are outdated and rendered especially useless because of the conflated roles of test developers and users under NCLB. It is no longer tenable to argue that test developers and/or users are responsible for investigating test impact. Instead, NCLB circumstances dictate a new reality where negotiated responsibility is paramount. A recently introduced framework (Nichols and Williams 2008) argues that test developers' responsibilities vary given the dynamics of *Breadth of Construct* (narrowly- or broadly-operationalised), *Test Score Use* (intended and unintended purposes), and *Time* (passage of time since an unintended use of the test became popular). The framework and arguments presented call for the increasing importance of upfront negotiations to determine how impact research will be undertaken and by whom.

Finally, the paper urges L2 researchers to incorporate, in addition to the reactive approach to documenting the educational and social influence of policies and test practices, anticipatory impact research, intended to improve policy making. The section introduces SIA, which underscores the importance of proactive research that can be utilised to inform and guide policy makers when formulating educational reform policies. SIA calls for systematic plans of investigations, the need to address different stakeholders, the importance of documenting intended and unintended changes, the focus on anticipating negative impact and offering suggestions for mitigation, and the criticality of working with policy makers.

References

Alderson, J C and Wall, D (1993) Does washback exist? *Applied Linguistics* 14, 115–29.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999) *Standards for educational and psychological testing*, Washington, DC: American Psychological Association.

Bailey, A L and Butler, F A (2003) *An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document*, (CSE Report 611), Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A L, Butler, F A and Sato, E (2005) *Standards-to-standards linkage under Title III: Exploring common language demands in ELD and science standards*, (CSE Report 667), Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Barrow, C J (2000) *Social impact assessment: An introduction*, Oxford, UK: Oxford University Press.

Burdge, R J (2007) www.socialimpactassessment.net/. Downloaded January 7, 2009.

Burdge, R J, Fricke, P, Finsterbursch, K, Freudenburg, W R, Gramling, R, Holden, A, Llewellyn, L G, Petterson, J S, Thompson, J and Wiliams, G (1995) Guidelines and principles for social impact assessment: Interorganizational Committee on Guidelines and Principles for Social Impact Assessment, *Environmental Impact Assessment Review* 15, 11–43.

Butler, F A and Stevens, R (2001) Standardized assessment of the content knowledge of English language learners K-12: Current trends and old dilemmas, *Language Testing* 18, 409–427.

Chalhoub-Deville, M (2008a) Assessments for K-12 English language learners: What tests are ELLs taking and ways to improve them, paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Chalhoub-Deville, M (2008b) *Standards-Based Assessment in the US: Social and Educational Impact*, plenary paper presented at the 3rd international conference of the Association of Language Testers in Europe, University of Cambridge, Cambridge: England.

Chalhoub-Deville, M (2009) The intersection of test impact, validation, and educational reform policy, *Annual Review of Applied Linguistics* 29, 118–131.

Chalhoub-Deville, M and Deville, C (2006) Old, borrowed, and new thoughts in second language testing, in Brennan, R L (Ed.) *Educational measurement* (4th ed.,), Washington, DC: The National Council on Measurement in Education & the American Council on Education, 516–30.

Chalhoub-Deville, M and Deville, C (2008) National standardized English language assessments, in Spolsky, B and Hults, F (Eds), *Handbook of Educational Linguistics*, Oxford, UK: Blackwell Publishers, 510–522.

Cheng, L (2005) *Changing language teaching through language testing: A washback study*, Studies in Language Testing 21, Cambridge: UCLES/ Cambridge University Press.

Cheng, L (2008) Washback, impact and consequences, in Shohamy, E and Horberger, N H (Eds), *Encyclopedia of language and education, Vol. 7: Language testing and assessment* (2nd ed.), Dordrecht, The Netherlands: Springer, 349–64.

Cummins, J (1981) Age on arrival and immigrant second language learning in Canada: A reassessment, *Applied Linguistics* 2, 132–149.

Cummins, J (2001) Assessment options for bilingual learners, in Tinajero, J V and Hurley, S R (Eds) *Literacy Assessment of Bilingual Learners*, Boston: Allyn and Bacon, 115–129.

Fradd, S H, McGee, P L and Wilen, D K (1994) *Instructional Assessment: An Integrative Approach to Evaluating Student Performance*, Reading, MA: Addison Wesley.

Francis, D J, Rivera, M, Lesaux, N, Kieffer, M, and Rivera, H (2006) *Practical guidelines for the education of English language learners: Research based recommendations for instruction and academic interventions*. Available at www.centeroninstruction.org/files/ELL1-Interventions.pdf.

Freeman, D and Freeman, Y (2001) *Between worlds: Access to second language acquisition*, Heinemann: Portsmouth, NH.

García, P (2003) The Use of High School Exit Examinations in Four Southwestern States, *Bilingual Research Journal* 27, 431–450.

Green, A B (2007) *IELTS washback in context: Preparation for academic writing in higher education*, Studies in Language Testing 25, Cambridge: UCLES/ Cambridge University Press.

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

Hawkey, R (2006) *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*, Studies in Language Testing 24, Cambridge: UCLES/ Cambridge University Press.

Herrera, S G and Murry, K G (2005) *Mastering ESL and bilingual methods: Differentiated instruction for culturally and linguistically diverse (CLD) students*, NY: Pearson.

Hulbert, A (May 2007) Standardizing the standards, *New York Times Magazine,* 11–12.

Kieffer, M J, Lesaux, N K and Snow, C E (2007) Promises and pitfalls: Implications of No Child Left Behind for identifying, assessing, and educating English language learners, in Sunderman, G (Ed.) *Holding NCLB Accountable: Achieving Accountability, Equity, and School Reform*, Thousand Oaks, CA: Corwin Press, 57–74.

Menken, K (2006) Teaching to the test: How No Child Left Behind impacts language policy, curriculum, and instruction for English language learners, *Bilingual Research Journal* 30, 521–546.

Menken, K (2008) *English learners left behind: Standardized testing as language policy*, Buffalo, NY: Multilingual Matters LTD.

Messick, S (1989a) Validity, in Linn R L (Ed.) *Educational measurement* (3rd ed.), Washington, DC: The American Council on Education & the National Council on Measurement in Education, 13–103.

Nichols, P and Williams, N (2008) *Evidence of test score use in validity: roles and responsibility*, paper presented at the annual meeting of the National Council on Measurement in Education, New York.

No Child Left Behind (2001) Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425.

Rivera, C (2008) *Testing accommodations for ELLs: What we know and what we still need to know*, paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Shohamy, E (1996) Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing* 13, 340–49.

Shohamy, E (2001) *The power of tests: A critical perspective on the uses of language tests*, Essex, England: Longman.
Spolsky, B (1997) The ethics of gatekeeping tests: What have we learned in a hundred years? *Language Testing* 14, 242–47.
Teachers of English to Speakers of Other Languages (TESOL) 1997: *ESL Standards for PreK-12 Students*, Alexandria, VA: Author.
Teachers of English to Speakers of Other Languages (TESOL) 2006: *Prek-12 English Language Proficiency Standards*, Alexandria, VA: TESOL Inc.
Wall, D (1996) Introducing new tests into traditional systems: Insights from general education and from innovation theory, *Language Testing* 13, 334–57.
Wall, D (2005) *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*, Studies in Language Testing 22, Cambridge: Cambridge University Press.
Wall, D and Alderson, J C (1993) Examining washback: The Sri Lankan impact study, *Language Testing* 10, 41–69.

17 The impact of large-scale and classroom-based language assessments on the individual

James E Purpura
Teachers College, Columbia University

Abstract

This paper explores how the introduction and use of assessments impact individuals, especially as this relates to individuals engaged in the teaching and learning process. I will begin by examining how the research evidence on test impact can be contextualised within a test validity argument for justifying test use. I will then explore the research on how large-scale, standardised tests – those external to the classroom context – impact individuals in the teaching and learning process. Finally, in trying to examine the interface between assessment and second language acquisition, I will discuss language processing from a cognitive-interactionist perspective and will highlight the role that assessment plays in SLA processes. I will use this learning-oriented model of assessment as a springboard for discussing the potential impact that assessment could have on individual learning in classroom contexts.

Introduction

A widely held belief in the introduction and use of language assessments is that the decisions we make on the basis of test score interpretations can have a significant impact on individuals, groups, institutions and society-at-large (Bachman 1990, Frederiksen and Collins 1989, Messick 1989, Popham 1987). The introduction and use of these assessments can also affect the processes and practices that stakeholders engage in (e.g. their mental and behavioural actions related to learning, teaching, parenting, evaluating), the products that are generated (e.g. score reports, test preparation materials), the policies they create and enforce (e.g. grading, standard-setting, admissions), and the enhancement of competencies over time (Alderson and Wall 1993, Bailey 1996, Brindley 2001, Hughes 1993, Wall 2005). Furthermore, given the range of decisions that could be made based on score interpretations (e.g. selection,

placement, progress, promotion, certification), the introduction and use of an assessment can have significant attitudinal, behavioural and developmental consequences for stakeholders (Bachman and Palmer 1996, Messick 1989). Impact and consequences are further heightened when large-scale, standardised assessments are perceived as gate-keeping or gate-opening devices associated with prestige, socio-economic mobility and meritocratic beliefs about an individual's access to education and employment (Alderson and Hamp-Lyons 1996, Bachman and Purpura 2008, Ross 2008). Finally, the introduction and use of assessments obviously entail intended and unintended consequences, which benefit some stakeholders, while simultaneously having a potentially adverse effect on others (Alderson and Hamp-Lyons 1996, Andrews 1995, Bachman 1990, Mandaus 1988, Messick 1989, 1996, Popham 1987, Shohamy 1993, Wall and Alderson 1993).

The research on the impact of language tests on individuals has been examined through many lenses. On the macro level, one strand of research focuses on complex issues of language test impact related to the wider socio-political contexts (Brindley 2001, McNamara and Roever 2006). This research has been motivated by forces in the social, economic, political and ethical arena and the concomitant use of language assessments as a strategy for implementing governmental policy (Hawthorne 1997, Menken 2008, National Research Council 1999, Shohamy and McNamara 2009). In this regard, several researchers (e.g. McNamara 2008, Shohamy and McNamara 2009) have examined the impact and consequences of the governmental use of assessments to address concerns about immigration, citizenship, asylum and the integration of immigrants into host societies. In this same vein, other researchers (e.g. National Research Council 1999) have raised questions and concerns about the impact and consequences of the government's ability to use high-stakes assessments of school subjects (delivered in English) to make fair and non-discriminatory classifications of abilities for students whose native language is not English, particularly when these assessments have not been validated for use with non-native speakers. They also question, on ethical grounds, the use of these classifications as a means of assigning or withholding educational services (e.g. ESL instruction) to individuals in need (e.g. English language learners).

Until now, the vast majority of theoretical and empirical research on language test impact has revolved around the impact and consequences of using assessments as a means of educational accountability and reform, typically in ways intended by institutional or governmental policymakers. In a climate of increased social, political and economic demands related to ongoing demographic changes, increased globalisation and the need to enhance the competencies of individuals in society, assessments have been used as change agents for raising learning (and teaching) standards, holding students (and teachers) accountable for gains in a broader range of competencies than in the

past, and for driving instruction and influencing curricula in ways intended by policymakers (Brindley 2001, Chalhoub-Deville and Deville 2008, Cheng 2004, 2008, Choi 2008, Norton 1997, Shohamy 1993, 2001, Qi 2005). In other words, assessments have been used to push reform – and implicitly create a *de facto* language policy (Menken 2008). For example, Qi (2005) described how the intended purpose of introducing the *National Matriculation English Test* in China was intended to move English language teaching from a 'linguistic' approach to a 'language use' approach. However, this strategy generally failed because, among other reasons, the test continued to use a multiple-choice format rather than a communicative language use one, which would have been more compatible with the test developers' intentions. Also, since the primary use of the test results was for university selection (and teacher and school evaluation), teachers felt compelled to 'teach to the test' rather than rethink their teaching methodology.

Since much of the research on test impact in applied linguistics has focused on how the introduction and use of standardised language tests affect general educational processes and stakeholder categorisations, those involved in second or foreign language classrooms might wonder how this research might speak to the introduction and use of language assessments in instructional contexts, where the primary purpose of assessment is not to make selection or certification decisions, but to collect trustworthy, score-based and descriptive information about what students know and can do with the language, so that this information can be used to inform and support educational decision-making. In other words, how can score-based inferences from classroom-based assessments be used to make decisions about student readiness to benefit from instruction, decisions about student attainment and growth, or decisions about the kind of feedback to provide learners at different points in the learning process? Similarly, how can teachers use this evidence to make decisions about next steps related to curricular content, instructional methods and classroom materials? In the context of language learning, then, we might ask: what impact evidence do we have about the intended consequences of our classroom-based assessments for individual learners and teachers, and to what extent do these assessments serve to promote or inhibit further learning and more effective teaching?

In this paper, I will explore how the introduction and use of assessments impact individuals, especially as this relates to individuals engaged in the teaching and learning process. I will begin with an examination of how test impact and consequences research fits into broader frameworks of test validation. I will then summarise and explore the research on how tests external to the classroom context appear to impact teaching and learning. Finally, in trying to examine the interface between assessment and second language acquisition, I will discuss language processing and the potential role that assessment plays in second language acquisition (SLA). I will use this as a

springboard for discussing the potential impact that assessment might have on individual learning in classroom contexts.

Test impact as evidence in a validity argument

When assessments are used to provide score-based decisions about individuals, these decisions, regardless of the stakes, will have consequences for the test takers. Since these consequences can potentially have an adverse effect on test takers, language testers need to justify their use to stakeholders. This can be done by formulating two kinds of arguments (Kane 2001, Mislevy, Steinberg and Almond 2002, 2003). First, by making an interpretative argument, we can specify inferences and assumptions related to the meaningful and appropriate interpretation of test scores. For example, we can answer the question: what is the evidence that a classroom-based test of writing ability will produce scores that are meaningful and appropriate for the construct being measured? In addition to the interpretive argument, we can make a validity argument, which allows us to evaluate the theoretical rationale and empirical evidence used to support claims and refute counter-claims about test use (Bachman 2005, Bachman and Palmer forthcoming, Chapelle, Enright and Jamieson 2008, Kane 1992, 2004, Messick 1989, 1996). In other words, given all the evidence related to this assessment, can the score-based interpretations about the students' knowledge, abilities or skills be used to make decisions, and are these decisions justified?

While the provision of test impact evidence is important, Messick (1989, 1996) argued that testers first need to establish that the properties of the assessment can provide adequate and appropriate interpretations of test score use and decision-making. In other words, we must offer theoretical and empirical evidence to support inferential claims about language test use (Messick 1989, 1996, Bachman and Palmer 1996, forthcoming). Messick described several claims that could be made about test use. One claim refers to the notion that the test content is relevant and adequately representative of the abilities we hope to measure. For example, if we are giving a speaking test at the B1 level, the test should reflect competencies typical of B1, and no other level of the *Common European Framework of Reference* (Council of Europe 2001). A second test use claim involves the overall structure of the test and the degree to which it can be theoretically and empirically justified as a measure of the knowledge, abilities and skills that we wish to measure. Third, we need to provide evidence to support claims that the test tasks engage not only the examinees' language abilities, but also their socio-cognitive and affective processes in construct-appropriate ways (Embretson 1983, cited in Messick 1996). For example, if our assessment goal is to measure our students' abilities to make a complaint, then the test tasks should engage examinees' abilities in ways similar to those invoked in

real-life task completion. A fourth type of claim refers to the notion that the interpretations, made on the basis of test scores, can generalise impartially across groups, settings and tasks. In other words, we need to ensure that our assessments are bias-free – for example, that the scoring of speaking tasks is consistent and equitable across test takers and test-taker groups. Last but not least, we need to show that the use of the test, along with its resulting decisions, will support the intended, beneficial consequences we hope to achieve, and we will need to document the unintended, detrimental consequences of test use. In this regard, Messick (1989, 1996) was particularly concerned with potential adverse consequences for individuals that stem from problems in the test itself, which may then lead to mistaken score-based interpretations and misguided actions. He wrote, 'any negative impact on individuals or groups should not derive from any source of test invalidity such as construct underrepresentation or construct-irrelevant variance' (1996:252). For example, if teachers emphasise a communicative approach in class, but their tests consist of only multiple-choice grammar questions, then we might say that something critical in the test construct is missing. As a result, students might place undue importance on learning how to answer multiple-choice grammar questions, rather than learning how to communicate in the target language.

In sum, whether we are designing a high or low-stakes test or justifying the introduction or use of an assessment, we need to bear in mind, along with other test properties, the potential beneficial effects of using an assessment for decision-making purposes. We also need to consider any potential unintended, negative consequences of these decisions.

Large-scale test impact on teaching and learning

The vast majority of theoretical and empirical research on test impact in applied linguistics has focused on the effects of assessments on teaching and learning, referred to in the literature as *washback* (Hughes 1989). Over the years, researchers have offered many definitions and depictions of washback. They have also studied the intended/unintended, positive/negative, and strong/weak effects that large-scale, externally mandated assessments have had on classroom teaching and learning.

Alderson and Wall (1993:117) defined 'washback' as 'the degree to which an assessment causes language teachers and language students to do things they would not necessarily otherwise do', arguing persuasively that washback must not be assumed, but investigated for specific areas of impact (i.e. the content of teaching, teaching methodology, materials, assessment practices), for the direction of washback (i.e. positive, negative), and for the extent of impact (i.e. strong, weak). In this regard, they posited 15 hypotheses related to what aspects of teaching and learning will be affected by tests.

Of these, several addressed how tests putatively impact individual test takers and teachers. These include:

Impact on individual test takers

- A test will influence learning.
- A test will influence how students learn.
- A test will influence what learners learn.
- A test will influence the rate and sequence of learning.
- A test will influence the degree and depth of learning.
- A test will influence attitudes to the content, method, etc. of teaching and learning.

Impact on individual teachers

- A test will influence teaching.
- A test will influence what teachers teach.
- A test will influence how teachers teach; and therefore [how students learn].
- A test will influence the rate and sequence of teaching.
- A test will influence the degree and depth of teaching.
- A test will influence attitudes to the content, method, etc. of teaching and learning.

(Alderson and Wall 1993:120–121)

In this study Alderson and Wall (1993) concluded that washback involves a complex array of interacting variables that mitigate against any firm predictions regarding the effects of test use in any given situation at any one time.

Bachman and Palmer (1996) viewed 'washback' in terms of the broader notions of test impact and consequences. For them impact operates at the micro level, where assessment influences individual test takers and teachers, and at the macro level, where it affects educational systems and society-at-large. In describing the effect of washback on individual test takers, they identified three ways in which examinees can be affected. First, examinees can be affected by the experience of taking and/or preparing for the exam. Second, the feedback they receive about their performance can impact their learning. Finally, they are affected by the decisions made about them on the basis of the test score interpretations.

Whether these effects are intended/unintended, positive/negative, or strong/weak for individual examinees depends, of course, on an examination of evidence. In some instances, tests can have a strong positive effect on learning by changing individual examinees' attitudes towards the test (i.e. greater value for the material taught) and by promoting construct-relevant actions (i.e. increased effort and learning) (Black, Harrison, Lee, Marshall and Wiliam 2003). This appears to be especially true when: (1) tests that are

aligned with curricular content that teachers support (Messick 1996, Pearson 1988, Resnick and Resnick 1992); (2) test tasks assess the kinds of behaviours that teachers value (Bailey 1996, Heyneman and Ransom 1990, Hughes 1989, Kellaghan and Grearney 1992, Wall 1996); and (3) tests provide score reports that supply information for further learning (Bailey 1996, Black et al 2003, Kellaghan and Grearney 1992). In other instances, tests can have a strong negative influence on learning, especially when there is no correspondence between assessment and instruction or when the primary goal of assessment is to succeed (for the purposes of selection or certification), rather than to learn. In most instances, however, washback effects are extremely complex and variable for different individuals in terms of the washback intent, direction (positive/negative) and intensity (strong/weak) (Green 2007).

At the micro level, Bachman and Palmer (1996) also described individual teachers, and indirectly instructional programmes, as being impacted by tests. From one perspective, the intended, beneficial washback effect of large-scale assessment use on classroom teachers is to provide normative information about the students' abilities and indirectly about instruction. Assessments have also been used to encourage teachers to reflect on their practice for the purpose of updating both the content of their curriculum and their methodology (e.g. Cheng 2008, Choi 2008, Qi 2005, Qian 2008, Ramanathan 2008, Wall 2005). Ideally, the washback effect of such assessments would result in a reprioritisation of curricular content, methods, materials, perceptions and attitudes. However, a considerable amount of research has shown that the washback effects of large-scale assessments are not so straightforward. Wall (1996) found that in studying the effects of a *New National Exam* on EFL instruction in Sri Lanka, the exam had a considerable impact on the content of instruction, the types of materials teachers used, the focus of instruction at different times, and ways in which assessments were designed, but it had little to no impact on the teaching methodology or on the scoring of performance. More recently, Perrone (2008) examined the longitudinal impact of the *First Certificate in English Exam* (FCE) (Cambridge ESOL) on teaching in a general EFL course compared with an FCE preparation course. He found that while the language skills and exam-related activities (e.g. timed practice tests, test-taking strategies, FCE content) generally became more test-like (especially in the FCE prep class) as the exam date neared, the students' mean scores on the FCE in both class types were not significantly different. As a result, the exam-related methods seemed to have had a very limited impact on student performance on the FCE. In fact, students in the general EFL course actually outperformed those in the FCE prep course to some degree.

In sum, many attempts to use externally mandated, high-stakes assessments to reform teaching and learning have not only had mixed results, but more seriously, some have incurred unforeseen negative consequences that have severely undermined the intended assessment goals (e.g. Alderson

and Hamp-Lyons 1996, Cheng 2005, Qi 2005, Shohamy, Donitsa-Schmidt and Ferman 1996, Wall 2005, Watanabe 1996). For example, Choi (2008) described the situation in Korea in which Koreans view the ability to communicate in English as the *sine qua non* for individual success in life; they also accept the function of assessment as a fair indicator of English language ability. However, stakeholders seem to have prioritised the goal of achieving a high score on the standardised language test over the true goal of learning to communicate in English. As a result, teachers and students spent valuable time learning how to answer multiple-choice test questions rather than learning how to communicate.

In such situations, Bachman and Palmer (1996) described how teachers might find themselves 'teaching to the test' rather than teaching the curriculum they prefer. This narrowing of the curriculum (Smith 1991) can be responsible for a reduced emphasis on skills that teachers perceive as time-consuming or complex (e.g. problem solving) (Frederiksen 1984), or it can result in a loss of actual teaching time in favour of more time for teaching test-taking strategies (Smith, Edelsky, Draper, Rottenberg and Cherland 1989). More seriously, exam pressures from high-stakes tests may have a deleterious effect on teacher attitudes toward teaching, testing and the educational system in general, especially if 'teaching to the test' is in conflict with what a teacher believes to be the best instructional choice for student learning.

In all of these situations, the empirical research has shown that the washback effects of large-scale assessments can have varying degrees of success. These effects can be caused by several interacting factors, which may potentially produce unintended, negative consequences to stakeholders. How these tests specifically affect individuals is not clear.

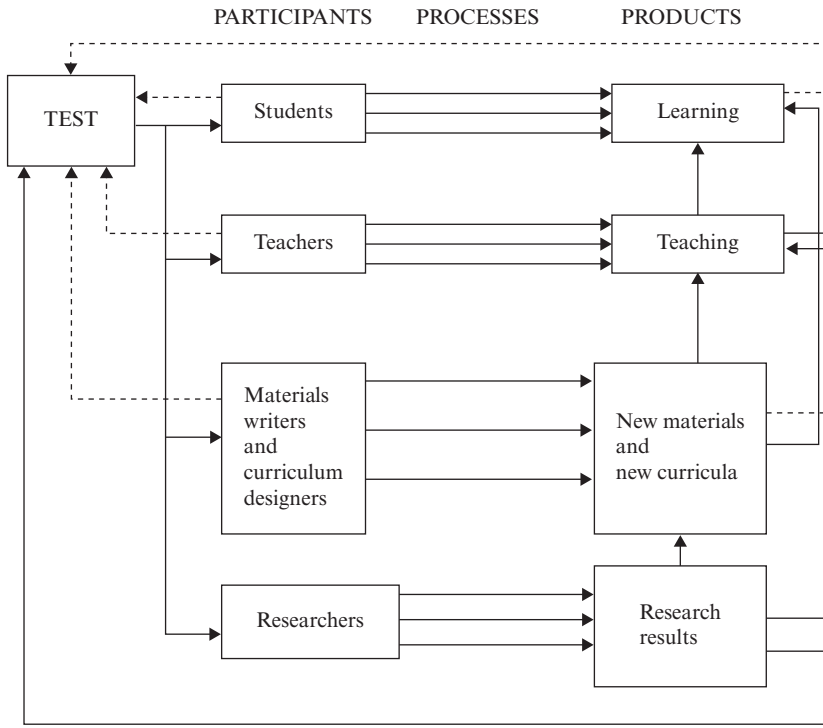
Bailey (1999) defined 'washback' as the extent to which an exam positively or negatively fosters or hinders the attainment of educational goals held by test stakeholders. Drawing on Hughes' (1993 as cited in Bailey 1996) influential trichotomy model of washback, she proposed a coherent model of washback in which she specified the interrelationships that result from the ways in which a test interacts with the participants (whose perceptions and attitudes may be affected by the test and its use), the processes (i.e. the actions taken by the participants to promote or understand teaching and learning), and the products associated with teaching and learning (i.e. what is learned and the depth of learning). This model is presented in Figure 1.

Using this model to examine the washback effects on students, Bailey (1996:264–65) identified 10 ways in which learners have been affected by tests.

1. Practising items similar in format to those on the test.
2. Studying vocabulary and grammar rules.
3. Participating in interactive language use.
4. Reading widely in the target language.

5. Listening to non-interactive language (radio, television, etc.).
6. Applying test-taking strategies.
7. Enrolling in test-preparation courses.
8. Requesting guidance in their studying and feedback on their performance.
9. Enrolling in, requesting or demanding additional (unscheduled) test-preparation classes or tutorials (in addition to or in lieu of other language classes.
10. Skipping language classes to study for the test.

Figure 1 Bailey's Basic Model of Washback (Bailey 1996:264)



In examining washback in Japan, Watanabe (1996) sought to discover the extent to which teachers incorporated translation into classroom practice as a result of a large-scale Japanese exam. Rather than finding a clear, linear relationship between the exam and teachers' actions, he reported considerable variability in the teachers' use of translation in class – and this seemed to depend on their personal beliefs, educational background and past learning

experiences. Later, exploring this complexity in more depth, Watanabe (2004) identified the following contributing factors to washback effects: dimension (specificity, intensity, length, intentionality, and the washback value), aspects of teaching and learning amenable to exam influence, and factors mediating the washback effect (prestige factors, macro-context factors). In describing 'dimension', he made particular reference to the *specificity* of the washback effect in terms of one or multiple components being affected, the *intensity* of the effect as it related to all individuals or only some, the *length* of the effect with regard to its short or long-term duration, the *intentionality* of the effect or the degree to which the consequences were intended or unintended, and the positive and negative *value* of the effect.

Endeavouring to disentangle the complexity of washback effects in academic writing, Green (2007:17) examined the impact of the IELTS on teaching and learning processes in both IELTS-prep courses and other courses designed to develop academic literacy skills. To examine this, he put forth a model of washback that relates the washback variability of examinees (who came with their own personal characteristics and values) to washback direction (positive or negative effect) and washback intensity (strong or weak). With regard to washback direction, he maintained that:

> Participants set the test stakes according to their awareness [. . .] of the uses to be made of test results. The stakes associated with the test influence the behaviour of those preparing for the test; high stakes encouraging greater adjustment on the part of participants. They also affect test design issues as higher test stakes impose stricter attention to questions of test fairness and encourage techniques that support objectivity.

Green (2007:24–25) also claimed that washback intensity (strong or weak effect) 'varies in relation to the participants' perceptions of the test stakes [. . .] and the test difficulty'. Finally, he hypothesised that 'washback will be most intense where participants:

- value success on the test above developing skills for the target language use domain
- consider success on the test to be challenging (by being both attainable and amenable to preparation)
- work in a context where these perceptions are shared (or dictated) by other participants.

In sum, much research has been carried out on how high-stakes tests putatively affect teachers and learners. This research has examined the positive and negative consequences of using government-mandated tests for decisions about immigration, citizenship, asylum and the integration of immigrants into host societies (Shohamy and McNamara 2009). It has also investigated

how the introduction and use of high-stakes, national tests (e.g. Ross 2008) or international tests can impact teaching and learning with mixed results. Some studies (Rea-Dickins 2004, Wall 2005, Wall and Alderson 1993) have even looked at the effects of standardised tests on teacher assessment practices.

Despite these advances, much remains to be learned, in my opinion, about effects that both large-scale and classroom-based assessments have on SLA. While some researchers (e.g. Alderson and Wall 1993, Andrews 1995, Andrews, Fullilove and Wong 2002, Bachman and Palmer 1996, Bailey 1996, Shohamy 1992) have examined the general effects that tests can have on learning, and while others have highlighted the importance and potential effects of using test results to provide learners with diagnostic information (Alderson 2005, Bachman and Palmer 1996, Lantoff and Poehner 2008, Shohamy 1994), I know of no studies that have seriously looked at the effects of how assessment might be *purposefully* used by teachers to engage individual learners in SLA processes. In other words, what role does assessment play in how learners process new information, how do they develop the ability to use this new information, or how do they benefit from feedback so as to develop deeper understandings of the new learning points? These questions are critical for a deeper understanding of how large-scale and classroom assessment impact individuals. In the next section, I will discuss the role and potential impact of assessment use in the learning process.

Learning-oriented language assessment: the role and potential impact of assessment on language learning in classrooms

Several researchers (Bachman and Cohen 1998, Purpura 2004, 2006, 2007, Rea-Dickins 2001, 2004, 2008, Rea-Dickins and Gardner 2000, Shohamy 1994) have expressed the view that in order to understand assessment practices and the potential they have for learning, testers need to explore the interfaces between language assessment and SLA research. Cheng (2008) and Wall (2000) also articulated the need to study the impact of assessment use on learning in classroom contexts. In order to explore how tests can impact individuals in learning another language, I would like to turn our attention away from the effect of tests on teaching and learning in large-scale contexts to the effect of assessment in classroom contexts, where, assessment, in one form or another, is central to the learning process.

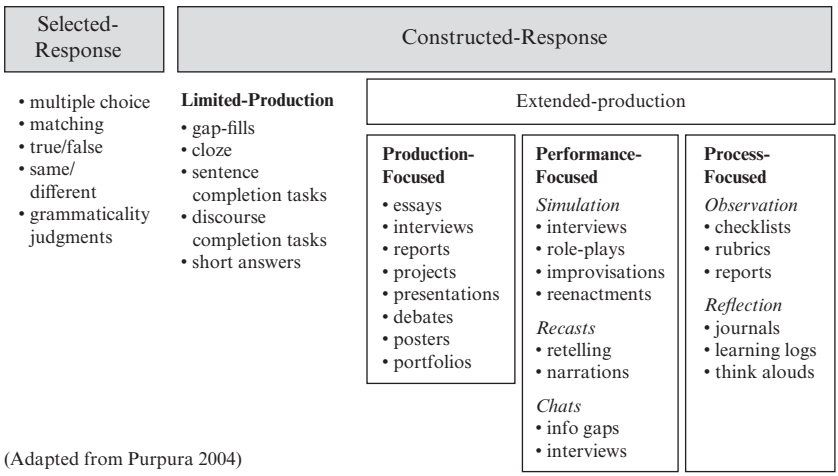
Language teachers have always understood that one goal of classroom assessment is to obtain information about how much students have learned in relation to a set of curriculum standards or objectives. The 'how much', or the evaluative part of the assessment, is motivated by the need to provide achievement results to different stakeholders. This normative evaluation of

performance is what Black and Wiliam (1998) refer to as 'assessment OF learning'. This is what most of the studies described so far have been about. An example of 'assessment OF learning' in the classroom context is the final exam. Such an assessment is seen as a culminating experience of learner mastery. And while most teachers give these tests assuming that some benefit will be gained by preparing for or taking the test, the focus is on the collection of information, so that interpretations can be made about attainment for the purpose of giving a grade. In most cases, the assessment information is not used to guide and support further learning with a concrete plan for feedback action on the part of teachers and learners.

Besides the need to provide assessments OF learning, language teachers have also understood that another aim of classroom assessment is to obtain information about what learners have processed at any particular point in time, so that they can determine to what extent this represents a gap in their knowledge or their ability to use this knowledge to communicate in another language. The information derived from more *informal*, often observation-based, assessments serves to make decisions about instructional next steps. In these instances, assessment is not an event per se; it is a pursuit of information with no *a priori* value judgment on task type. In fact, to obtain meaningful information for pedagogical decision-making, teachers employ a wide range of assessment strategies at critical moments during the course of the lesson. I call these 'testable moments' since they play a central role in how learners process information and develop competence. Figure 2 presents an overview of common task types used in classroom-based assessment.

The majority of 'testable moments' in teaching are, in fact, not associated

Figure 2 Task types – Expected response formats



(Adapted from Purpura 2004)

with test forms or grades; they involve the gathering of information through observation, the confirmation of comprehension, the identification of learning gaps, and the provision of information related to the current and desired performance. This is usually followed by some intervention strategy (what I call 'teaching') and the collection of further information to confirm or disconfirm the impact of instruction on individual learning (what I include in 'assessment'). This approach to classroom-based assessment acknowledges the synergistic and recursive links among curriculum, teaching, learning and assessment. Strongly rooted to a model of SLA, this approach to classroom-based assessment is what I have called a 'learning-oriented approach to assessment' (Purpura 2004, 2006, 2007). In this approach, the individual learners and what they need to close learning gaps on the route to acquisition are at the core of assessment. And all activities undertaken by teachers and students have the explicitly, intended consequence of providing information that can inform decisions about how to guide and support learning. This is also what Black and Wiliam (1998) refer to as 'assessment FOR learning' or 'assessment THROUGH learning'. I refer to this simply as 'learning-oriented language assessment', as I feel, 'learning' should be prioritised, rather than assessment and its medium.

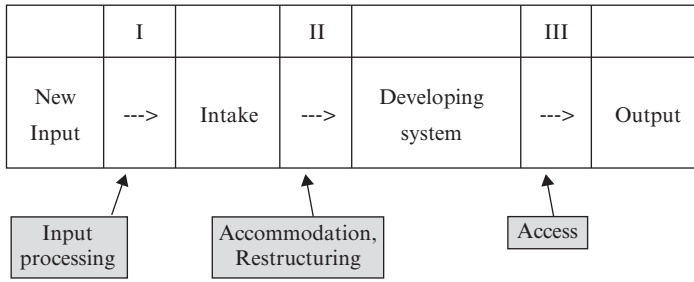
In considering whether classroom-based assessment has any effect on learning processes and products, I believe that constructs measured by means of assessments need to be informed *not only* by a model of language proficiency or performance, *but also* by some explicit model of second or foreign language learning (Purpura 2004, 2006, 2007), whether that be for characterising the knowledge, abilities and skills learners have gained thus far or for providing guidance and support for further learning (Pellegrino, Baxter and Glaser 1999).

In this section, I will examine how assessment fits into the broader notion of learning. This arises from my belief that the implicit, intended purpose of all classroom-based assessments, and arguably all large-scale assessments, should be to inform and direct individual learning in some way. I will first describe, from an SLA perspective, how learners process new learning targets. I will then discuss the role that assessment can play in this process. If assessment is *ever* to have an impact on individuals, it is in the process of sustained learning that we will discover this. Finally, as relatively little research in language testing has looked at the role of assessment use on learning processes in classroom contexts (see Rea-Dickins and Gardner 2000, Rea-Dickins 2001, 2004, 2008), learning-oriented language assessment can, I believe, serve as an organising frame for research on the interface between assessment and learning.

Learning-oriented language assessment

In order to understand how learning-oriented assessment works, I will briefly describe a model of second language processing. I will then discuss

Figure 3 VanPatten's (1996) depiction of three sets of processes in second language acquisition and use (p. 154)



how assessment practices might be embedded in this model. In so doing, I will highlight potential 'testable' moments as they relate to this process. Throughout the discussion, I will explain the potential impact that assessment use might have on individual learning. Before beginning, I want to mention that while other approaches to SLA exist, I take a socio-cognitive or cognitive-interactionist view of SLA in this discussion.

VanPatten (1996) conceptualised the three basic processes in second language acquisition and use as the result of internal, cognitive mechanisms that allow learners to map linguistic forms onto their meanings by means of input, and input processing, so that these forms and meanings can ultimately be accessed to produce output. As seen in Figure 3, VanPatten posits three processes of SLA and use: (1) input processing; (2) accommodation and restructuring – or the integration of new linguistic data into the developing system; and (3) access – or the retrieval of the new linguistic data for production. In his research, VanPatten focuses mainly on the first process – input processing.

To illustrate the full model, a learner is first presented with new input such as a past tense verb form. Noticing something different about the input, the learner begins to process it by making a connection between the form and its meaning. If the past tense form is noticed (see Schmidt's noticing hypothesis, 1990, 1995) and associated with a completed action in the past, the result of this form-meaning mapping is 'intake'. VanPatten (1996:7) argues convincingly that since 'acquisition is intake-dependent' and 'intake is in turn input-dependent', then, input processing is a critical stage in SLA that, in my opinion, presents an important opportunity for informal (but also for formal) assessment. At the intake stage of the process, the mapping may or may not be totally complete or accurate. However, once intake has occurred, the new form is available for further processing, and may be accommodated into the learner's developing linguistic system. This may result in a restructuring of the learner's linguistic system to include past tense forms as a means of expressing past time.

As the learner may have thus far acquired *only* an understanding of the new learning point, several researchers (e.g. DeKeyser 1997, 1998, 2007, Pienemann 1998, Swain 1985) have argued that the learner must then engage in repeated practice if he wishes to internalise the new feature so that it can be accessed and retrieved automatically. The importance of practice, as a means of building representations in long-term memory to overcome the limitations of working memory, is also highlighted in theories of skill acquisition (Anderson 2000). While a discussion of the issues surrounding practice is beyond the scope of this paper (see DeKeyser 2007 for an insightful treatment of the topic), practice – especially when accompanied by feedback – is also critical to SLA.

Practice, in my opinion, involves a principled set of 'classroom-based assessment' tasks (often called 'pedagogical tasks') designed to elicit learning target(s) so that learners will have opportunities to: (1) deepen their knowledge of the learning point, (2) develop the ability to use the learning target meaningfully in context, (3) develop skill in using the learning target automatically in interaction, and (4) receive feedback that promotes reflection and steps for improvement. As mentioned above, no *a priori* value judgments should be placed on task type provided they meet the assessment purpose. Assessment in this context is designed to make explicit to individuals their knowledge, abilities and skills, with the intention that conscious reflection and further learning will be fostered (and monitored). These assessments are meant to be descriptive, not evaluative.

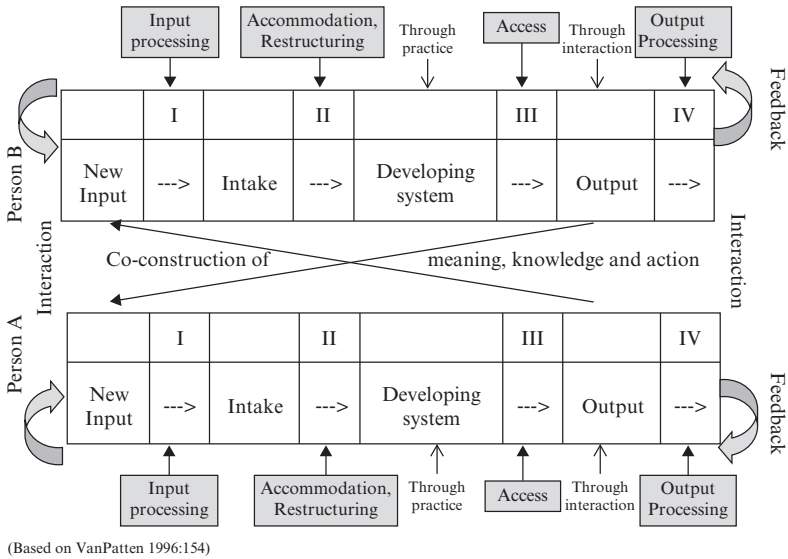
While practice is important, practice needs to be accompanied by corrective feedback for learning to occur (National Research Council 2001). In fact, probably the most important feature of classroom assessment is the provision of descriptive (not evaluative) feedback to learners in the hope that they will use this information to reflect on the learning gap and ultimately, choose to close it. In fact, empirical research in SLA has consistently shown that the provision of feedback seems to have a substantial positive effect on learning – even though we have no conclusive research evidence on how and when to give feedback (Leeman 2007). Empirical research in mainstream education has also shown that learning gains can be significantly increased when the feedback includes:

- data on the actual level of some measurable attribute
- data on the desirable level of that attribute
- a mechanism for comparing the two levels and assessing the gap between them
- a mechanism by which the information can be used to alter the gap (Black et al 2003).

In a seminal study on feedback, Kluger and DeNisi (1996) found that learning gains increased dramatically when learners were not only given information about their performance – that is, mastery of the learning goals, but were also told how to use this information to improve.

With regard to the kind of practice needed for SLA, several researchers (e.g. de Bot 1996, Muranoi 2007, Robinson 2001, Skehan 1998, Swain 1985, 2005) have strongly argued that practice must provide learners with the opportunity to produce meaningful output. They have shown that output, viewed not just as the 'production of forms and meanings at the sentential level' but as 'interaction with others' (VanPatten 2004:27) pushes learners to notice new language, formulate and test hypotheses about language, reflect consciously on their language successes and failures (through language repair mechanisms), and develop fluency and automaticity. To characterise this perspective in a way that includes both the cognitive and social dimensions of learning, consider a cognitive-interactionist model of SLA, as seen in Figure 4.

Figure 4 Cognitive–interactionist model of SLA (New)



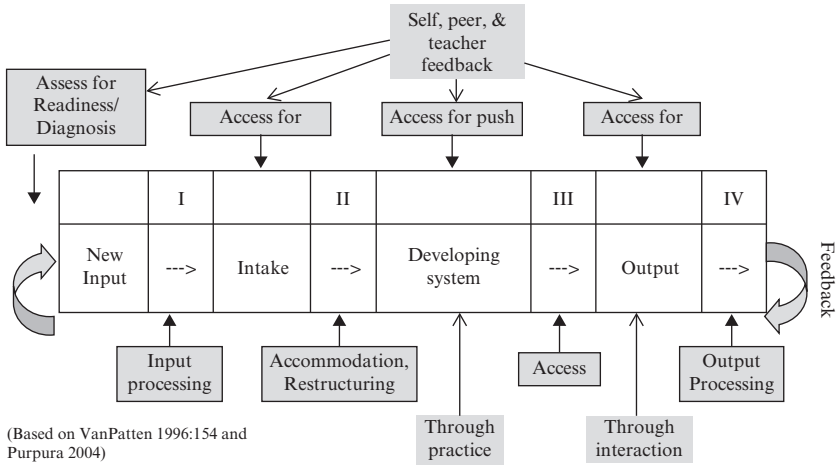
(Based on VanPatten 1996:154)

As the learning target is proceduralised through practice and feedback, the speed and accuracy of using the new form to communicate increases in a 'systematic, non-linear pattern over successive attempts' (National Research Council 2001:85). Also, given the typological similarities between the L1 and the L2 and a number of other factors, the processing of new learning targets may be quick at first, but then followed by 'subsequent and continuous improvements in performance that accrue at a slower and slower rate' (National Research Council 2001:85).

In short, the learning process obviously offers multiple opportunities for assessment that can significantly impact learning. Figure 5 shows how

assessment could be brought into the learning process for individuals and how feedback can play a role.

Figure 5 Learning–oriented model of language assessment



The first opportunity for learning-oriented assessment is prior to instruction. In classroom contexts, a teacher might want to know her students' strengths and weaknesses in narrative writing prior to planning instruction. To get this information, she administers a diagnostic test that requires students to write a narrative. This allows her to identify targets of learning. The intended goal of this writing diagnostic is to acquire an understanding of the learners' strengths and weaknesses so that instruction can be designed to close the learning gaps. I call this 'assessment for diagnosis'. The intended impact of these assessments is to better match instruction to the learners' individual needs. For further information on diagnostic assessment, see Alderson (2005), Edelenbos and Kukanek-German (2004), Hughes (1993) and Huhta (2008).

SLA researchers commonly use pre-instruction assessments to determine if learners are 'developmentally ready' for instruction (tests of readiness). Developmental readiness here refers to learners' level of attainment along the interlanguage continuum, the assumption being that a learner acquires certain linguistic features in a fixed order and sequence of acquisition. For example, if an SLA researcher would like to study the acquisition of genitive relative clauses ('whose book I borrowed'), they might design a test to determine if the learners are 'developmentally ready' to learn this structure.

Another opportunity for learning-oriented assessment is post instruction – or after learners have 'theoretically' processed the new learning target by incorporating it into their linguistic system or when they have 'theoretically'

developed the ability to use the new feature in meaningful interaction. From an SLA perspective, this could be referred to as 'assessing for output'. As seen in Figure 1, there are many ways of assessing for output depending on the goal of assessment. For example, a 'structure-output task' (e.g. in a discourse completion task) could be designed to elicit the production of forms and meanings at the sentence or discourse levels (see Lee and VanPatten 2003). Similarly, a performance-based task designed to elicit sustained interaction of some sort (e.g. a simulation) could be devised. The general goal of assessing for output is to provide information on mastery. This information could then be used to supply meaningful, descriptive feedback to learners and teachers. One intended, beneficial consequence of this type of assessment for learners is to learn how their performance compared to the expected performance, so that they can engage in the process of focused repair and the closing of learning gaps. While many teachers might view this as an assessment OF learning, it could equally be treated as an assessment FOR learning.

Many other 'testable moments' occur during the teaching/learning process. When a learner has processed new linguistic input, teachers usually 'test for intake'. This is often referred to as a 'comprehension check'. Given the role of intake for SLA processing (Ellis 1997, VanPatten 1996), I would argue that 'assessing for intake' should be planned for with much greater care than a spontaneously devised comprehension check. Many of the selected-response methods would probably be appropriate for these assessments. For examples of intake tasks, see Lee and VanPatten (2003) or Purpura (2004), or see Ellis's (1997) interpretation tasks.

Finally, during the course of instruction, teachers have multiple opportunities to provide learners with assessment-like activities designed to push the developing system through practice of all sorts. What is critical to remember is that practice alone does not necessarily lead to acquisition; feedback is essential. A number of language testing researchers (e.g. Cheng and Warren 2005, Matabonga, Kenyon and Carpenter 2005, Patri 2002) have begun to examine the feedback provided in the context of self- and peer-assessment. SLA researchers have also done extensive research on the feedback conditions that contribute to intake and that push the developing system toward full acquisition (see Ellis 2008).

Conclusion

The purpose of this paper was to explore the impact of large-scale and classroom-based tests on the individual. We have seen that the research base for understanding the impact of large-scale assessments on individuals is complex and context-dependent. In some instances, the introduction and use of an exam is perceived as having a positive effect on individuals, while in other cases, exam use can have serious, unintended negative consequences. Aside

from the socio-political dimension, most of the impact research was concerned with the impact that high-stakes, standardised exams had on teaching and teaching programmes. Again in some instances, the introduction and use of an exam helped initiate teacher reflections about their practice, resulting in the desired change, whereas in other instances, the introduction and use of a new exam had a limited effect or simply failed to incur the desired change – and in the process, created another set of problems for individual teachers and learners. Unfortunately, this research has very little to say about the effect of these exams on learning processes. What is the nature of socio-cognitive engagement when preparing for an exam – even cramming for an exam? To what extent does the content of these exams and the assessment method illuminate learning gaps for teachers and learners? What do teachers and learners do to close these gaps? Finally, what are the socio-cognitive opportunity costs of preparing for standardised exams that we feel are not aligned with constructs and methods we value?

Another aim of this paper was to explore the relationship between classroom-based assessment and individual learning in classroom contexts. The vast majority of empirical research on this topic has been concerned with the teacher as an agent of assessment (Rea-Dickins 2004). When it comes to formative assessment research in language classrooms or the impact of classroom-based assessments on learning, the language assessment research has only recently begun to provide some insights. Bachman and Palmer (1996:98) contrast the formative and summative functions of information derived from language tests, stating that formative information could 'help students guide their own subsequent learning' or it could '[help] teachers modify their teaching methods and materials so as to make them more appropriate for their students' needs, interests and capabilities'. Bachman (1990:55) also mentions how feedback from tests can potentially improve teaching and learning outcomes as well as educational processes. In the first serious empirical study on formative assessment in language testing, Rea-Dickins and Gardner (2000) examined classroom assessment practices in nine inner-city schools. They concluded that despite perceptions, the decisions made on the basis of assessments were high-stakes given the serious consequences for children of a false positive or false negative classification. Like Teasdale and Leung (2000), they question the assessment criteria used in standardised assessments as appropriate criteria for classroom assessment. While several researchers have recently highlighted the need to focus on the relationship between classroom-based assessment and learning, none have proposed that the discussion begin with an examination of the learning process in light of how assessment might be used for different purposes with different individuals on the path to second or foreign language acquisition. If we really wish to examine the impact of assessment on individuals, I believe we need to begin with learning. Also important is how teachers are capable

of using assessments and assessment information to guide and support this process. In describing the learning-oriented model of language assessment, I have attempted to bring an initial sense of coherence to the topic. Obviously, research now needs to follow.

References

Alderson, J C (2005) *Diagnosing foreign language proficiency: the interface between learning and assessment*, London: Continuum.

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

Alderson, J C and Wall, D (1993) Does washback exist? *Applied Linguistics* 14, 116–29.

Anderson, J (2000) *Cognitive psychology and its implications* (5th ed.), New York: Worth.

Andrews, S (1995) Washback or washout? The relationship between examination, reform and curriculum innovation, in Nunan, D, Berry, V, and Berry R, (Eds) *Bringing about change in language education*, Hong Kong: University of Hong Kong 67–81.

Andrews, S, Fullilove, J and Wong, Y (2002) Targeting washback – A case study, *System* 30, 207–225.

Bachman, L F (1990) *Fundamental considerations in language testing*, Oxford: Oxford University Press.

Bachman, L F (2005) Building and supporting a case for test use, *Language Assessment Quarterly: An international Journal* 2 (1), 1–34.

Bachman, L F and Cohen, A D (1998) Language testing – SLA interfaces: an update, in Bachman, L F and Cohen, A D (Eds) *Interfaces between second language acquisition and language testing research*, Cambridge: Cambridge University Press, 1–31.

Bachman, L F and Palmer, A S (1996) *Language assessment in the real world: Developing language assessments and justifying their use*, Oxford: Oxford University Press.

Bachman, L F and Palmer, A S (Forthcoming) *Language assessment in practice,* Oxford: Oxford University Press.

Bachman, L F and Purpura, J E (2008) Language assessments: Gate-keepers or door-Openers? in Spolsky, B and Hult, F M (Eds) *The Handbook of Educational Linguistics*, Oxford, UK: Blackwell Publishing, 456–468.

Bailey, K M (1996) Working for washback: A review of the washback concept in language testing, *Language Testing* 13, 257–279.

Bailey, K (1999) *TOEFL Monograph Series, MS15: Washback in language testing,* Princeton: Educational Testing Service.

Black, P, Harrison, C, Lee, C, Marshall, B and Wiliam, D (2003) *Assessment for learning: Putting it into practice*, New York: Open University Press.

Black, P and William, D (1998) Assessment and classroom learning: *Assessment in Education*, 5 (1), 7–71.

Brindley, G (2001) Outcomes-based assessment in practice: Some examples and emerging insights, *Language Testing* 18 (4), 393–408.

Brown, J D (1997) Do tests washback on language classrooms? *The TESOLANZ Journal* 5, 63–80.

Chalhoub-Deville, M and Deville, C (2008) Nationally mandated testing for

accountability: English language learners, in Spolsky, B and Hult, F M (Eds), *The Handbook of Educational Linguistics* 510–522.
Chapelle, C, Enright, M and Jamieson, J (2008) Test score interpretation and use, in Chapelle, C, Enright, M, Jamieson, J (Eds) *Building a validity argument for the Test of English as a Foreign Language,* New York: Routledge 1–25.
Cheng, L (2004) The washback effect of a public examination change on teachers' perceptions toward their classroom teaching, in Cheng, L and Watanabe, Y (Eds) with Curtis, A (Ed.) *Washback in Language Testing: Research Contexts and Methods*, Mahwah, NJ: Lawrence Erlbaum Associates, 147–170.
Cheng, L (2005) *Changing language teaching through language testing: A Washback study*, Cambridge: Cambridge University Press.
Cheng, L (2008) The key to success: English language testing in China, *Language Testing,* 25 (1), 15–37.
Cheng, W and Warren, M (2005) Peer assessment of language proficiency, *Language Testing,* 22 (1), 93–121.
Choi, I-C (2008) The impact of testing on EFL education in Korea, *Language Testing,* 25 (1), 39–62.
Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.
de Bot, K (1996) The psycholinguistics of the output hypothesis, *Language learning,* 46 (3), 529–55.
DeKeyser, R (1997) Beyond explicit rule learning: Automatizing second language morphosyntax, *Studies in Second Language Acquisition* 19, 105–221.
DeKeyser, R (1998) Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar, in Doughty, C and Williams, J (Eds), *Focus on form in classroom second language acquisition*, Cambridge: Cambridge University Press, 42–63.
DeKeyser, R (Ed.) (2007) *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*, Cambridge: Cambridge University Press.
Edelenbos, P and Kukanek-German, A (2004) Teacher assessment: the concept of "diagnostic competence", *Language Testing* 21 (3), 259–83.
Ellis, R (1997) *SLA research and language teaching*, Oxford: Oxford University Press.
Ellis, R (2008) *The study of second language acquisition*, Oxford: Oxford University Press.
Frederiksen, N (1984) The real test bias: Influences of testing on teaching and learning, *American Psychologist* 29 (3), 193–202.
Frederiksen, J R and Collins, A (1989) A systems approach to educational testing, *Educational Researcher* 18 (9), 27–32.
Green, A (2007) *IELTS washback in context: Preparation for academic writing in higher education*, Cambridge: Cambridge University Press.
Hawthorne, L (1997) The political dimension of English language testing in Australia, *Language Testing* 14 (3), 248–260.
Heyneman, S P and Ransom, A W (1990) Using examinations and testing to improve educational quality, *Educational Policy* 4 (3), 177–192.
Hughes, A (1989) *Testing for language teachers*, Cambridge: Cambridge University Press.
Hughes, A (1993) *Testing for language teachers* (2nd Edition), Cambridge: Cambridge University Press.

Huhta, A (2008) Diagnostic and formative assessment, in Spolsky, B and Hult, F M (Eds), *The Handbook of Educational Linguistics*, Oxford, UK: Blackwell Publishing, 469–482.

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

Kane, M T (2001) Current concerns in validity theory, *Journal of Educational Measurement* 21 (1), 31–35.

Kane, M (2004) Certification testing as an illustration of argument-based validation, *Measurement* 2, 135–170.

Kellaghan, T and Grearney, V (1992) *Using examinations to improve education: A study of fourteen African countries*, Washington, DC: The World Bank.

Kluger, A N and DeNisi, A (1996) The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory, *Psychological Bulletin* 119 (2), 254–284.

Lantoff, J P and Poehner, M E (2008) Dynamic assessment, in Shohamy, E and Hornberger, N (Eds) *Encyclopedia of Language and Education, 2nd Edition, Volume 7: Language Testing and Assessment*, New York: Springer, 273–284.

Lee, J F and VanPatten, B (2003) *Making communicative language teaching happen* (2nd Edition), Boston, MA: McGraw Hill.

Leeman, J (2007) Feedback in L2 learning: Responding to errors during practice, in DeKeyser, R (Ed.) *Practice in a second language: Perspectives from applied linguistic and cognitive psychology*, Cambridge: Cambridge University Press, 111–37.

Mandaus, G F (1988) The influence of testing on the curriculum, in Tanner, L N (Ed.) *Critical issues in Curriculum*, Chicago, Illinois: Chicago University Press, 83–121.

Matabonga, V, Kenyon, D and Carpenter, H (2005) Self-assessment, preparation, and response time on a computerized oral proficiency test, *Language Testing* 22 (1), 59–92.

McNamara, T (2008) The socio-political and power dimensions of tests, in Shohamy, E and Hornberger, N (Eds) *Encyclopedia of Language and Education, 2nd Edition, Volume 7: Language Testing and Assessment*, New York: Springer 415–427.

McNamara, T and Roever, C (2006) *Language testing: The Social dimension*, Malden, MA: Blackwell.

Menken, K (2008) High-stakes tests as de facto language education policies, in Shohamy, E and Hornberger, N (Eds) *Encyclopedia of Language and Education, 2nd Edition, Volume 7: Language Testing and Assessment*, New York: Springer 401–413.

Messick, S (1989) Validity, in Linn, R I (Ed.) *Educational Measurement* (3rd ed.), New York; Macmillan 113–103.

Messick, S (1996) Validity and Washback, *Language Testing* 13 (3), 241–256.

Mislevy, R J, Steinberg, L S and Almond, R G (2002) Design and analysis in task-based language assessment, *Language Testing* 19, 477–496.

Mislevy, R J, Steinberg, L S and Almond, R G (2003) On the structure of educational assessments, *Measurement: Interdisciplinary Research and Perspectives* 1, 3–62.

Muranoi, H (2007) Output practice in the L2 classroom, in DeKeyser, R (Ed.) *Practice in a second language: Perspectives from applied linguistic and cognitive psychology*, Cambridge: Cambridge University Press, 51–84.

National Research Council (1999) *High stakes: Testing for tracking promotion and graduation.* Committee on Appropriate Test Use, Heubert, J P and Hauser, R M (Eds) Commission on Behavioral and Social Sciences and Education, Washington, DC: National Academy Press.

National Research Council (2001) *Knowing what students know: The science and design of educational assessment*, Washington, D.C.: National Academy Press. The Committee on the foundations of assessment, Pellegrino, J W, Chudowsky, N and Glaser, R (Eds) Bureau of Testing and Assessment.

Norton, B (1997) Accountability in language assessment, In C Clapham and D Corson (Eds.). *Encyclopedia of Language and Education, Vol. 7: Language Testing and Assessment* (313–322), Dordrecht: Kluwer Academic publishers.

Patri, M (2002) The influence of peer feedback on self-and peer-assessment of oral skills, *Language Testing* 19 (2), 109–132.

Pearson, (1988) Tests as levers of change, in Chamberlain, D and Baumgartner, R (Eds) *ESP in the classroom: Practice and evaluation*, Oxford: Modern English Publications, 98–107.

Pellegrino, J W, Baxter, G P and Glaser, R (1999) Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice, in Iran-Nejad, A and Pearson, P D (Eds) *Review of research in education (Volume 24)*, Washington, DC: American Educational Research Association, 307–353.

Perrone, M (2008) *The impact of the First Certificate in English (FCE) examination upon the EFL classroom: A washback study*, unpublished Doctoral Qualifying Paper 6500B, Teachers College, Columbia University, New York.

Pienemann, M (1998) *Language processing and second language development; Processability theory*, Philadelphia: John Benjamins.

Popham, J (1987) Two decades of educational objectives, *International Journal of Educational Research* 11 (1) 31–41.

Purpura, J E (2004) *Assessing grammar*, Cambridge: Cambridge University Press.

Purpura, J E (2006) *Issues and challenges in measuring SLA*, Colloquium presentation: Towards theoretically meaningful L2 assessments for SLA research, American Association for Applied Linguistics Conference, June 2006.

Purpura, J E (2007) Conceptualizing and measuring meaning in SLA Research, presented at the Second Language Research Forum, University of Illinois, Champaign-Urbana (Keynote Address).

Qi, L (2005) Stakeholders' conflicting aims undermine the washback functions of a high-stakes test, *Language Testing* 22 (2), 142–73.

Qian, D (2008) English language assessment in Hong Kong: A survey of practices, developments and issues, *Language Testing* 25 (1), 85–110.

Ramanathan (2008) Testing of English in India: A developing concept, *Language Testing* 25 (1), 111–126.

Rea-Dickins, P (2001) Mirror, mirror on the wall: Identifying processes of classroom assessment, *Language Testing* 18 (4), 429–462.

Rea-Dickins, P (2004) Understanding teachers as agents of assessment, *Language Testing* 21 (31), 249–258.

Rea-Dickins, P (2008) Classroom-based assessment, in Shohamy, E and Hornberger, N (Eds) *Encyclopedia of Language and Education,*

2nd Edition, Volume 7: Language Testing and Assessment, New York: Springer, 257–271.
Rea-Dickins, P and Gardner, S (2000) Snares and sliver bullets: Disentangling the construct of formative assessment, *Language Testing, 17*(2), 215–243.
Resnick, L B and Resnick, D (1992) Assessing the thinking curriculum: New tools for educational freeform, in Gifford, B G and O'Conner, M D (Eds) *Changing assessments: Alternative views of aptitude, achievement and instruction*, Boston, MA: Kluwer Academic Publishers, 37–75.
Robinson, P (2001) Task complexity, cognitive resources, and syllabus design: A triadic theory of task influences on SLA, in Robinson, P (Ed.) *Cognition and second language instruction*, New York: Cambridge University Press, 284–318.
Ross, S (2008) Language testing in Asia: Evolution, innovation and policy challenges, *Language Testing* 25 (1), 5–13.
Schmidt, R W (1990) The role of consciousness in second language learning, *Applied Linguistics* 11, 129–158
Schmidt, R W (1995) Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning, in Schmidt, R W (Ed.) *Attention and awareness in foreign language learning*, Honolulu: University of Hawaii Press, 1–63.
Shohamy, E (1992) Beyond proficiency testing: A diagnostic feedback testing model for assessing language learning, *The Modern Language Journal*, 76, 513–21.
Shohamy, E (1993) The power of tests: The impact of language tests on teaching and learning, NFLC Occasional Paper, College Park, MD: National foreign Language Center, University of Maryland.
Shohamy, E (1994) The role of language tests in the construction of second-language acquisition theories, in Tarone, E, Gass, S and Cohen, A (Eds) *Research Methodology in Second Language Acquisition*, Hillsdale, NJ: Lawrence Erlbaum, 133–142.
Shohamy, E (2001) *The Power of Tests: A critical Perspective on the Uses of Language Tests*, Longman/Pearson Education: London.
Shohamy, E and McNamara, T (2009) Editorial: Language tests for citizenship, immigration and asylum, *Language Assessment Quarterly*, 6 (1), 1–5.
Shohamy, E, Donitsa-Schmidt, S and Ferman, I (1996) Test impact revisited: A study of washback, *Language Testing* 13 (3), 298–317.
Skehan, P (1998) *A cognitive approach to language learning*, Cambridge: Cambridge University Press.
Smith, M L (1991) Put to the test: The effects of external testing on teachers, *Educational Researcher* 20 (5), 8–11.
Smith, M L, Edelsky, C, Draper, K, Rottenberg, C and Cherland, M (1989) *The role of testing in elementary schools*, Center for Research on Educational Standards and Student Tests, Graduate School of Education, UCLA, Los Angeles, CA.
Swain, M (1985) Communicative competence: Some roles of comprehensible input and comprehensible output in its development, in Gass, S and Madden, C (Eds) *Input and second language acquisition*, Rowley, MA: Newbury House, 235–253.
Swain, M (2005) The output hypothesis: Theory and research, in Hinkel, E (Ed.) *Handbook of second language teaching and learning*, Mahwah, NJ: Lawrence Erlbaum, 471–81.

Teasdale, A and Leung, C (2000) Teacher assessment and psychometric theory: A case of paradigm crossing? *Language Testing* 17 (2), 163–184.
VanPatten, B (1996) *Input processing and grammar instruction in second language acquisition*, Boston: McGraw Hill.
VanPatten, B (2004) *Processing instruction: Theory, research and commentary*, Mahwah, NJ: Lawrence Erlbaum Associates.
Wall, D (1996) Introducing new tests into traditional systems: Insights from general education and from innovation theory, *Language Testing* 13 (3), 334–352.
Wall, D (2000) The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System* 28, 499–509.
Wall, D (2005) *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*, Cambridge: Cambridge University Press.
Wall, D and Alderson, J C (1993) Examining Washback: The Sri Lankan impact study, *Language Testing* 10, 41–69.
Watanabe, Y (1996) Investigating washback in Japanese EFL classrooms: Problems and methodology, *Australian Review of Applied Linguistics* 13, 208–239.
Watanabe, Y (2004) Methodology in washback studies, in Cheng, L and Watanabe, Y (Eds) with Curtis, A (Ed.) *Washback in Language Testing: Research Contexts and Methods,* Mahwah, NJ: Lawrence Erlbaum Associates, 147–170.

18 A study of the Cambridge Proficiency in English (CPE) exam washback on textbooks in the context of Cambridge ESOL exam validation

Roger Hawkey
Consultant to Cambridge ESOL

Abstract

This paper, derived from a presentation with the same title at the ALTE Cambridge Conference in April 2008, reports an empirical study commissioned by Cambridge ESOL into the washback of the Certificate of Proficiency in English (CPE) on textbooks used on programmes preparing candidates for the exam. A key aim of the study was to produce evidence to be used in Cambridge ESOL's analysis of the validity of the CPE, in particular its *consequential* validity. In the study, each of a selection of 10 textbooks was evaluated independently by two experienced language-teaching specialists, using an adaptation of the textbook analysis instrument first developed and validated for the study of the impact of the International English Language Testing System (IELTS) exam. The evidence collected and analysed supported, as expected, a hypothesis that there is strong washback from the CPE exam to the textbooks in their treatment of language macro- and micro-skills, task types, language elements and topics. There was also evidence of changing washback on the textbooks as the CPE exam itself was revised. Important to an understanding of the role of exam-to-textbook washback in the exam provider's analysis of the consequential validity of the CPE (and related) exams are the study's finding that evaluators not only expect textbooks to represent the constructs, content, activities and tasks of an exam directly, but also to help develop learners' general language ability. This article draws conclusions from the study not only on exam-to-textbook washback but also on the role of exam washback and impact studies in Cambridge ESOL's research and validation policy.

Washback, impact and consequential validity

Cambridge ESOL (UCLES EFL as was) has long been concerned with the ways in which its exams for speakers of English as an alternative language affect a whole range of stakeholders. As long ago as 1943, the *Cambridge Examinations in English for Foreign Student Survey* asks 'how far examinations of this kind may act as a stimulus and a focusing point for both teachers and taught, and thereby promote the expansion of the studies which they are designed to test' (p. 37). Sixty-five years later, in the same month as the Cambridge 3rd international conference of the Association of Language Testers in Europe (ALTE) in April 2008, Hughes (2008:4), in an article in the UK *Weekly Guardian* presuming to address stakeholders involved both directly and less directly with Cambridge ESOL exams, insists that exams need 'to match the many varied needs of the students and requirements of employers and universities'.

Exams such as the Cambridge Proficiency in English (CPE) exam, at Common European Framework of Reference (CEFR) level C2, top of the range of the Cambridge ESOL Main Suite of general English exams, are considered by most of their stakeholders to have both *washback and impact*. These exams are *high-stakes*, that is 'seen, rightly or wrongly by students, teachers, administrators, parents, or the general public as being used to make important decisions that immediately and directly affect them' (Madaus 1988:87).

Of the four core themes of the ALTE 3rd International Conference, the source of papers in this volume, this account of the CPE Textbook Washback (CPETWB) study is relevant both to *language assessment for teaching and learning*, and *language assessment for stakeholder constituencies*. The study was carried out as part of Cambridge ESOL's systematic process for the validation of its exams, including their *consequential validity* (Messick 1989). Consequential validity is one of the validity types in a *socio-cognitive* framework for test validation adopted by Cambridge ESOL (influenced by Bachman 1990, Bachman and Palmer 1996, Chapelle, Jamieson and Enright 2004, Mislevy, Steinberg and Almond 2003 and Toulmin 2003). Messick's consequential validity construct is interpreted by Weir (2005), in a test validation framework developed with and used by Cambridge ESOL, as covering aspects of test validity such as impact on institutions and society, washback on individuals in the classroom or workplace, and avoidance of test bias. In addition to consequential validity, Weir's framework covers cognitive, context, scoring and criterion-related validities.

In the definitions which informed the study of the CPE's influences on textbooks, washback is taken to refer to an exam's influences on teaching,

teachers, learning, curriculum and materials (see Alderson and Wall 1993). The superordinate term impact is generally defined as 'the total effect of a test on the educational process and on the wider community' (McNamara 2000: 133). Washback is thus *part of* impact (see Green and Hawkey 2004, Hamp-Lyons 1998, McNamara 1996, 2000, Shohamy 2001). When Bachman and Palmer (1996:29) see impact as operating on two levels, a 'macro' level 'in terms of educational systems and society in general' and a micro level, 'a local and personal level, in terms of the people who are directly affected by tests and their results', they appear to be referring to a distinction similar to impact *vs* washback. Khalifa and Weir (2009) certainly consider that the consequential validation of an exam such as the CPE must cover both impact and washback.

So, the focus of the CPETWB study is on how evidence of exam washback relationships with textbooks strengthens or weakens Cambridge ESOL arguments on the *consequential validity* of its exams. The study must be seen as part of Cambridge ESOL's continuous and iterative search for evidence for the validity of its exams through routine analyses of test material, test takers' performance throughout the exam production, trialling, marking/grading, post-exam analyses and validation cycle (see, for example, Saville in Weir and Milanovic, Eds, 2003).

Washback research

Now, although washback may be narrower than impact, it is still extremely complex (as indicated by Alderson 2004, Bailey 1996, Cheng and Curtis 2004, Hawkey 2006b, Spratt 2005, Watanabe 2004). Alderson and Wall (1993) identify 15 washback hypotheses, involving the potential influence of a test on *the teacher and the learner, what and how teachers teach and learners learn, the rate and sequence of learning, and attitudes to teaching and learning methods*. These hypotheses indicate the complex of intervening variables likely to be operating between an exam and the way an exam-oriented textbook actually affects learners. Textbook washback evidence thus has to be probed carefully, interpreted seriously and sensitively in exam validation studies. We need to heed the warnings of researchers against too many clear-cut assumptions about it. Hamp-Lyons (2000) cautions against the over-simplification that exam washback necessarily leads to 'curricular alignment', Green and Hawkey (2004:66) that it will always be a 'harmful influence'. That teachers use methods indicated in exam syllabuses, teacher guides and textbooks because these will develop the skills required by the exam is questioned by Alderson and Wall (1993) and Wall (2005). Nor is washback unidirectional, i.e. from exam to textbook and teaching rather than bi-directional, i.e. also from textbook and teaching to exam (Wall 2005, CPE *Handbook*, UCLES 2002). Wall (2005) also raises empirical doubts that

teachers are driven by the exam rather than the textbook or that they overfocus on skills in the textbook that are tested in the exam (see also Hawkey 2006b). Nor are teachers' claims and their actual teaching necessarily the same (Wall 2005), or teacher and student perceptions of exam-preparation lesson content always similar (Hawkey 2006a).

Despite the importance of exam-to-textbook washback, however, Spratt (2005:8) in her survey of studies of exam washback on *curriculum, materials, teaching methods, feelings and attitudes, and learning* finds only Hamp-Lyons (1998), Shohamy, Donitsa-Schmidt and Ferman (1996), Lumley and Stoneham (2000) addressing teaching materials as a 'main issue'. Yet though the issue of *textbook* washback has indeed not been much addressed, we may add to Spratt's survey Cheng (1997), Saville and Hawkey (2004), Shohamy (1993) and Tsagari (2008). Cheng (1997:57), for example, finds that 'textbook publishers have certainly changed the forms and organisations of the teaching contents according to the new examination formats'. Note, too, how both the Cheng (1997) and the CPETWB study described below exemplify examination : textbook washback research in the interests of test validation, rather than for the more traditional purposes of textbook selection, by teachers and institutions, of the right textbook for their particular needs (see, for example, Sheldon 1988).

The CPE exam and the textbook washback (CPETWB) study

The CPE exam, first administered in 1913 and operating at CEFR C2 (Mastery) level (ALTE level 5), is top of the range in the Cambridge Main Suite of exams. The specification, validation and systematic revision of CPE in tune with the constructs of the language teaching and testing times are thus a matter of great importance to Cambridge ESOL. Weir and Milanovic (Eds) (2003), in a sister volume in this *Studies in Language Testing* series, chart the history of the CPE and detail the processes and products involved in updating the exam for 2002. Their book is recounted mainly in the words of the Cambridge ESOL subject officers working in the Assessment and Operations Group (AOG) on the exam papers involved. The AOG is responsible for managing and producing all Cambridge ESOL examinations and assessments, and provides operations functions that relate to examination administration (pretesting, entry and results processing, clerical and examiner marking and security of the examinations) and post-exam processing. The group also develops new tests and provides institutional versions of existing test products (Cambridge ESOL functional brief, April 2007).

Table 1 summarises the format and contents of the pre-revision and 2002 revised versions of the exam as presented in the respective exam *Handbooks.*

Table 1 Summary of the pre and post-revision versions of the CPE exam

Papers	CPE pre-revision	Revised CPE
1	Reading Comprehension: comprehension of gist, detail, tone, register; wide knowledge of vocabulary, usage, grammatical control	Reading: understand meaning of written English at word, phrase, sentence, paragraph, whole text level
2	Composition: write non-specialised descriptive, narrative, discursive texts; range of topics and tasks	Writing: write specified text types with range of functions
3	Use of English: knowledge, control of language system	Use of English: knowledge, control of language system
4	Listening Comprehension: extract information, interpret speakers' attitudes, recognise implications of stress, intonation	Listening: understand meaning of spoken English, extract information, understand speakers' attitudes, opinions
5	Interview: approx. 15 mins. Candidates tested individually, in pairs or groups of three 1: comparing thematically linked photos 2: comment on passage read silently 3: joint problem-solve activity	Speaking: 19 mins. Two candidates, two examiners: 1: interview, general interactional 2: collaborative task, visual, spoken prompts 3: individual long turns on written question; follow-up discussion

The aim of the Cambridge ESOL AOG, the Research &Validation group initiators of the CPETWB study and of the project co-ordinator (and writer of this paper) was to design a project to provide, for exam validation purposes, answers to research questions such as:

- How and to what extent has the CPE exam impacted on textbooks designed for use with CPE students?
- How is exam washback affected by changes in the CPE exam?
- What relationships are indicated between washback from the exam to the textbooks and the language proficiency needs of CPE exam takers?

The textbook washback study design

Given the perceived need for in-depth, expert-informed, qualitative data, the books selected for the CPETWB study were to be exam coursebooks rather than books of practice tests, as likely to be more revealing of the complex relationships between exam and textbooks in preparation class use. In line with the objectives and scale of the study, ten published CPE-oriented textbooks were selected, by professionals in AOG (see above), according to the following categories:

- four books aimed at the 1984 to June 2002 version of the exam
- four revised versions of those books aimed at the post-2002 CPE revision
- two completely new CPE-oriented textbooks.

Each of the 10 CPE-oriented textbooks was to be evaluated independently by two specialists selected for their experience with the exam. Nine of the 10 evaluators had taught on CPE and other Cambridge ESOL exam preparation courses, six were examiners, and four test item writers, three of these for the CPE exam itself. Four of the evaluators were working in the UK, two in Greece, and one each in Italy, Poland and Switzerland, variously at language schools or colleges, British Council centres, universities and a primary school. Evaluators received a commissioning letter informing them that the study was part of the continuing programme to validate, update and refine the CPE exam.

Each evaluator was to rate a pre-revision and a revised or new edition of a CPE book. Table 2 summarises the evaluator : book schema.

Table 2 Evaluator : textbook rating schema

Books	**Evaluators**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A1 (1 = pre-CPE revision edition)		x	x							
A2 (2 = post-CPE revision edition)		x	x							
B1				x			x			
B2				x			x			
C1					x					x
C2					x					x
D1								x	x	
D2								x	x	
E (new, post-CPE revision)	x					x				
F (new, post-CPE revision)	x					x				

The instrument for the analysis of test materials (IATM)

Given the aims and design of the CPETWB project above, an instrument was clearly required that would invite the collection of comprehensive and detailed information and opinion on the many areas of potential washback from exam to textbook. The Instrument for the Analysis of Test Materials (IATM) had been developed originally by Bonkowski (1996) and a Lancaster University team led by Charles Alderson, commissioned by Cambridge ESOL to facilitate IELTS impact study data collection. The original version of the IATM was trialled and validated between 1996 and 2000 (see Hawkey 2006b, Saville

and Hawkey 2004). The instrument performed satisfactorily in a study of the accessibility of IELTS General Training modules to 16 to 17 year old candidates (Smith 2004) and in the IELTS impact study (Hawkey 2006b). Given the relevance and generally positive experience using the IATM in washback studies, it was decided that the instrument should also be used, in an adapted form, in the CPETWB study. IATM was first adapted using the CPE exam specifications (pre- and post-revision versions, as detailed, for example in the 1984 *Changes of Syllabus in 1984* and 2003 CPE *Handbook* (UCLES). This adaptation was modified through two iterations according to feedback from senior team members of the Cambridge AOG (see above).

Given its test validation context, the approach of IATM is not so much to ask participants *whether* an exam influences textbooks; it is to be expected that it will. Rather, the IATM seeks to define relationships between the exam and the books so that they may be checked for positive consequential validity. This is done by asking users, in this case the 10 expert informants:

- which features of the CPE exam they perceive to be represented in the books
- how and to what extent these features are represented
- what they think of the treatments of the exam's features in the book(s).

The full IATM as used in the CPETWB study is appended. The instrument collects user information and opinion, yes/no, quantitative and qualitative, on the following elements of a textbook:

- *book type, units of organisation*
- *language features, enabling skills covered*
- *question / tasking techniques*
- *communicative genres, media, activities*
- *text topics, authenticity.*

Then finally three completely open-ended items, on:

- *the treatment of the four macro-skills*
- *the book as a whole*
- *the book's relationships with the target exam in general, in terms of its potential to help candidates with the CPE.*

Data validation

Although the IATM had been validated for the study of IELTS impact (see above), further post-validation evidence was sought from the modified IATM instrument used in the CPE study. With two independent evaluations of each textbook, instrument and inter-rater reliability evidence was obtained from a comparative analysis of pairs of evaluator responses. The following were key findings, according to the different data response types used, which had been selected using evidence from the trialling of the original IATM.

- On relatively clear-cut Yes/No questions such as *type of book,* raters achieved complete agreement (see below).
- On Yes/No items where options were limited to the presence or absence of a feature, for example a book's *units of organisation in terms of topics,* agreement across different raters of the same book was maximum or near-maximum.
- On items seeking responses on the presence and the prominence of categories, for example explicit practice of *language features*, where 18 selection options were listed, each with a three-choice quantity scale (e.g. a lot, a little, none), an average of 23 out of 36 identical selections were made across pairs of evaluators. For a complex construct such as *enabling (*or *micro-) skills* (again with 18 selection options) but with only presence or absence choices, pairs of raters of the same book averaged 27 out of 36 identical selections.

Figure 1 Comparison between two evaluators on enabling skills covered in one textbook

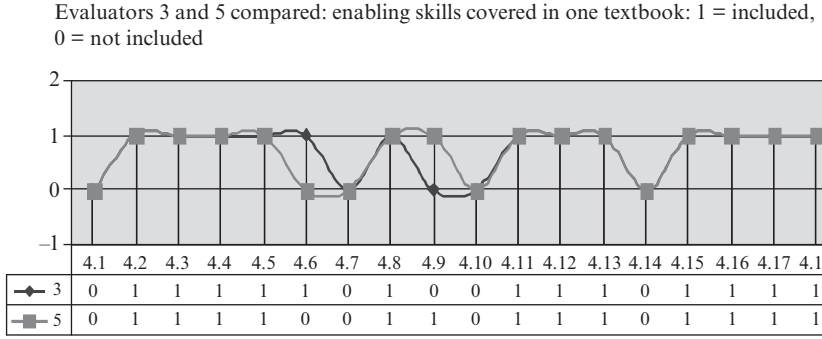


Figure 1 presents, as a representative example of the validation of evaluator responses for this study, the across-rater reliability analysis carried out on a pair of textbook evaluations. Where only one line in the comparative graph diagram is visible, as in the case here with 16 of the 18 micro-skills covered) there is agreement between the two evaluators. Graph lines seen separately suggest disagreement on the coverage of the enabling skill concerned (in this case 4.6 *evaluating evidence*, and 4.9 *distinguishing fact from opinion*; see Appendix for the full IATM).

Comparisons between responses to IATM items from evaluators of *different* books, to complete a form of convergent-discriminant validation on the items, revealed the expected weaker matches across ratings. The three exclusively open-ended comment invitations ending the IATM (see above) elicited data which provided both extra insights into test : textbook

washback, and triangulation evidence on item validity across closed and open data sources.

Summary findings

The full internal Cambridge ESOL report of the study (Hawkey 2004) includes 17 summary tables of evaluator comments in response to all IATM items, referenced to the particular books to which they were directed. Main findings on exam : textbook washback and likely to inform CPE consequential validation were as follows:

- Evaluators find that that specified CPE *topic range* and *skills* base are reflected positively in the textbooks (as in 13 positive (+) and 1 negative (−) comment).
- 20 of the evaluations (80%) state explicitly that the CPE communicative construct *should* be reflected in the textbooks.
- 33 comments (22+, 11−) concern whether the evaluators feel this match is satisfactorily achieved and/or that construct representation should be modified when the exam is revised.
- But evaluators want and expect the textbooks to fulfil *a teaching as well as a test-practice function*, meeting high *support* and *appearance* (64 comments, 35+, 29−) standards and with eight comments referring explicitly to standards on teaching–learning related criteria such as *organisation, grading/recycling, interest, teacher.*
- There are 15 open-ended overall comments from 20 evaluations (equally distributed between positive and negative) which remind that exam-oriented textbooks should allow for test takers *who are still below the target language proficiency level and need to be helped towards it;* the evaluators feel that the books should include 'skills-getting' and 'skills recycling/using in a meaningful manner' as appropriate to learners' proficiency level, needs and interests.

Note, returning to the running theme of the complexity of exam washback, the many variables intervening between exam and textbooks. The IATM seems quite good at getting at these complexities.

Conclusions

The purpose of this study is to add to consequential validity evidence for the Cambridge ESOL CPE exam, in particular the washback of the exam on preparation textbooks.

The following main conclusions may be drawn from the data analyses and discussion in this article:

1. The hypothesis that the pre-revision and revised CPE exams wash back on the evaluated textbooks is supported strongly. The 20 evaluations of the 10 textbooks, pre-revision, post-revision and new, all acknowledge the close relationship between the CPE exam and the books.
2. Given this clear finding, it is likely and, in fact, supported by the evidence of the study, that revisions in the CPE exam are marked by corresponding changes across editions of the textbooks.
3. All 20 evaluations imply that the books concerned *should* represent the text topics, enabling skills, activities and tasks of the exam. They also typically indicate aspects of the exam which they consider receive insufficient coverage in the books.
4. But the complex nature of washback identified in the literature review above is also reflected in the findings of the study. The evaluators expect the textbooks to 'teach as well as test'. They expect the books to take a learning-developmental approach, allowing for students who are not yet at the proficiency level (C2) of the CPE exam. There should be, the evaluators indicate, skills-getting and skills recycling and remedial work to help learners to improve their general language ability as appropriate to their needs and interests, as well as to help them succeed in the CPE exam.

This study is limited to one high-stakes exam. Similar studies of the washback of other exams on textbooks should also produce washback data of use in exam consequential validation. The textbook analysis instrument used here would seem amenable to adaptation for such further studies. It is possible that interview or focus group data (similar perhaps to those collected as part of the validation of the original IATM) would enhance the findings of future studies, the participants possibly to include textbook writers and teachers currently using the selected textbooks. Test : textbook washback studies could also play a useful part in classroom focused language exam and learner gain studies.

References

Alderson, C (2004) Foreword to Cheng, Watanabe and Curtis (Eds).

Alderson, C and Wall, D (1993) Does washback exist? *Applied Linguistics* 14, 115–129.

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

Bachman, L (2005) Building and supporting a case for test utilization, *Language Assessment Quarterly* 2 (1), 1–34.

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

Bailey, K (1996) *Washback in Language Testing*, Princeton New Jersey: Educational Testing Service.

Bonkowski, F (1996) *Instrument for the Assessment of Teaching Materials*, unpublished MA assignment, Lancaster University.
Chapelle, C, Jamieson, J and Enright, M (Eds) (2004) *Building a validity argument for the Test of English as a Foreign Language,* Mahwah, NJ: Laurence Erlbaum Associates.
Cheng, L (1997) How does washback influence teaching? Implications for Hong Kong, *Language and Education* 11, 38–54.
Cheng, L and Curtis, A (2004) Washback or Backwash: a review of the impact of testing on teaching and learning, in Cheng and Watanabe (Eds).
Cheng, L and Watanabe, Y (Eds) (2004) *Washback in Language Testing: Research Contexts and Methods,* New Jersey: Lawrence Erlbaum Associates.
Green, A and Hawkey, R (2004) Test washback and impact, *Modern English Teacher* 13 (4), 66–70.
Hamp-Lyons, L (1998) Ethical test preparation practice: the case of TOEFL, *TESOL Quarterly* 32 (2), 329–337.
Hamp-Lyons, L (2000) Social, professional and individual responsibility in language testing, *System* 28, 579–591.
Hawkey, R (2004) *Cambridge ESOL CPE Textbook Washback Study: Full report*, Cambridge: University of Cambridge ESOL Examinations.
Hawkey, R (2006a) Teacher and learner perceptions of language learning activities*, English Language Teaching Journal* 60 (3).
Hawkey, R (2006b) *Impact theory and practice: studies of the IELTS test and Progetto Lingue 2000,* Cambridge: Cambridge ESOL/Cambridge University Press.
Hughes, J (2008) Testing times for examiners, *The Guardian Weekly*, 7 March.
Kane, M, Crooks, T and Cohen, A (1999) Validating measures of performance, *Educational Measurement: Issues and Practice* 18 (2), 5–17.
Khalifa, H and Weir, C J (2009) *Examining Reading: Research and practice in assessing second language reading,* Cambridge: Cambridge ESOL/Cambridge University Press.
Lumley, T and Stoneham, R (2000) Conflicting perspectives on the role of test preparation in relation to learning? *Hong Kong Journal of Applied Linguistics* 5 (1), 50–80.
Madaus, G (1988) The influence of the testing on the curriculum, in Tanner, L (Ed.) Critical Issues in Curriculum: *87th yearbook of the National Society of Education*, Chicago: University of Chicago Press, 83–121.
McNamara, T (1996) *Measuring Second Language Performance*, Harlow: Longman.
McNamara, T (2000) *Language Testing*, Oxford, Oxford University Press.
Messick, S (1989) Validity, in Linn, R (Ed.) *Educational measurement*, New York: Macmillan, 12–103.
Mislevy, R, Steinberg, L and Almond, R (2003) On the structure of educational assessments, *Measurement*: *Interdisciplinary Research and Perspectives* 1 (1), 3–62.
Saville, N (2003) The process of test development and revision within UCLES/ EFL, in Weir and Milanovic (Eds) 2003.
Saville, N and Hawkey R (2004) The IELTS Impact Study: Investigating Washback on Teaching Materials, in Cheng and Watanabe (Eds).
Sheldon, L (1988) Evaluating ELT textbooks and materials, *ELT Journal* 42, 237–246.
Shohamy, E (1993) *The power of tests: The impact of language testing on teaching and learning,* Washington DC: The National Foreign Language Center.

Shohamy, E (2001) *The Power of Tests: a Critical Perspective on the Use of Language Tests,* Harlow: Pearson Education.

Shohamy, E, Donitsa-Schmidt, S and Ferman, I (1996) Test impact revisited: Washback effect over time, *Language Testing* 13, 298–317.

Smith, J (2004) IELTS Impact: a study on the accessibility of IELTS GT Modules to 16–17 year old candidates, *Research Notes* 18, 6–8.

Spratt, M (2005) Washback and the classroom: The implications for teaching and learning of studies of washback from exams, *Language Teaching Research* 9 (1), 5 –29.

Toulmin, P (2003) *The uses of argument,* Cambridge: Cambridge University Press.

Tsagari, D (2008) *Revisiting the concept of language washback: Results from an empirical study*, presentation given at the ALTE 3rd International Conference, 10–12 April 2008, Cambridge.

University of Cambridge Local Examinations Syndicate (1943) *Cambridge Examinations in English for Foreign Student Survey.*

University of Cambridge Local Examinations Syndicate (1982) *Changes of Syllabus in 1984.*

University of Cambridge Local Examinations Syndicate (1998) Certificate of Proficiency in English: *Handbook.*

University of Cambridge Local Examinations Syndicate (2002, 2003) Certificate of Proficiency in English: *Handbook.*

Wall, D (1997) Impact and washback in language testing, in Clapham, C (Ed.) *The Kluwer Encyclopaedia of Language in Education, vol 7. Testing and Assessment*, Kluwer, Dordrecht, 334–343.

Wall, D (2005) *The Impact of High-stakes Examinations on Classroom Teaching: A case study using insights from testing and innovation theory,* Cambridge: Cambridge ESOL/Cambridge University Press.

Watanabe, Y (2004) Methodology in Washback Studies, in Cheng and Watanabe (Eds).

Weir, C J (2005) *Language Testing and Validation: an evidence-based approach*, London: Palgrave Macmillan.

Weir, C J and Milanovic, M (Eds) (2003) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*. Cambridge: UCLES/Cambridge University Press.

Appendix

INSTRUMENT FOR THE ANALYSIS OF TEXBOOK MATERIALS (IATM)

0. The textbook being analysed:

Title			
Author(s)			
Publisher			
Place of publication		Year of publication	
Which students are you teaching using this book?			
What materials in addition to this book, if any, do you use when teaching the students?			

Questions on the kind of book this is, in general aim and organisation

1. What kind of book would you say this is? (Pease tick (✓) the box ❑ where appropriate.)

1.1 a language teaching book with no specific reference to international tests ❑	1.2 a book of practice tests only ❑	1.3 a language teaching book and an international test preparation book combined ❑
1.4 If it is a test-related book, for which test(s)?		
1.5 Any other comment on the type of book this is?		

If the book is a book of practice tests only, please go to Question 4.
If the book contains teaching material as well as practice tests, please go to Question 2.

2. The book's units / chapters etc. seem to be organised mainly according to: (Please tick (✓) the box(es) ❑ where appropriate; *more than one possible*.)

2.1 topics, themes ❑	2.2. language skills ❑	2.3 grammatical structures ❑	2.4 tests, tasks ❑	2.5 notions, functions ❑
2.6 other (please specify)				
2.7 Any further comment on the organisation of the book?				

Now a question on whether the book tries to break the target language down and teach the elements of the listening, reading, writing and speaking skills.

3. Your analysis of the book's explicit practice of <u>language features:</u> (Please tick (✓) appropriate spaces.)

	A lot	A little	None		A lot	A little	None		A lot	A little	None
3.1 recognition of sounds				3.2 pronunciation of sounds				3.3 stress and intonation			
3.4 grammar				3.5 sentence patterns				3.6 punctuation			
3.7 notions, functions				3.8 word formation				3.9 connotation			
3.10 synonymy				3.11 collocation				3.12 idioms			
3.13 linking words, expressions				3.14 paragraphs & discourse organisation				3.15 register			
3.16 other language components or features given explicit practice:											
3.17 related comments on how the book treats language and language features:											

Questions 4, 5 and 6 ask whether the book teaches and/or tests particular <u>enabling or micro-skills</u>, using a variety of techniques and activities.
Try checking 4, 5 <u>and</u> 6 before you write any comments, because skills, question / tasking and activities clearly overlap.

4. Enabling <u>skills</u> you think are covered in the book: (Please tick (✓) appropriate boxes ❑.)

4.1 understanding and conveying meaning through stress and intonation ❑	4.2. retrieving and stating factual information ❑	4.3 identifying main points ❑	4.4 identifying overall meaning ❑
4.5 predicting information ❑	4.6 evaluating evidence ❑	4.7 making inferences ❑	4.8 persuading, recommending ❑
4.9 distinguishing fact from opinion ❑	4.10 recognising roles ❑	4.11 identifying attitudes ❑	4.12 planning and organising information ❑

4.13 drawing conclusions ❑	4.14 narrating ❑	4.15 describing ❑	
4.16 Other skills covered by the book (please specify):			
4.17 Further comment on skills covered or not covered by the book:			

5. Your summary of the use of <u>question / tasking techniques</u> in the book: (Please tick (✓) appropriate spaces.)

	Frequent	A little	None		Frequent	A little	None
5.1 multiple / dual choice				5.2 true / false			
5.3 matching				5.4 gap filling / completion			
5.5 transformation				5.6 substitution			
5.7 paraphrasing				5.8 open-ended questions			
5.9 linking / joining				5.10 sequencing			
5.11 sentence / paragraph insertion				5.12 note taking / making			
5.13 summary				5.14 expansion			
5.15 correcting / editing				5.16 other techniques (please specify)			
5.17 Further comment on question and task techniques covered or not covered in this book							

6. Your evaluation of the extent to which the materials provide / encourage the following kinds of communicative opportunities. (Please tick (✓) in the appropriate spaces.)

	A lot	Quite a lot	Very little	None		A lot	Quite a lot	Very little	None
6.1 pair communication					6.2 group discussions and debates				
6.3 games, puzzles, quizzes					6.4 role play, simulations				
6.5 surveys, other project work					6.6 report writing				
6.7 review writing					6.8 essay writing				
6.9 letter writing					6.10 IT e.g. telephone, fax, email, Web use				
6.11 article writing					6.12 creative writing				
6.13 listening, reading, viewing for personal interest					6.14 other communicative opportunities *(please specify)*:				
6.15 Further comment on the communicative opportunities offered by the book:									

Questions 7 and 8 ask for information on test types and topics to check the coverage of the books.

7. How would you categorise the text types (heard, spoken, read, written) in the book? (Please tick (✓) appropriate boxes ❑.)

7.1 public announcement ❑	7.2 lecture/ talk ❑	7.3 radio/ TV report ❑	7.4 interview ❑
7.5 conversation ❑	7.6 discussion/ debate ❑	7.7 press report ❑	7.8 feature article ❑
7.9 correspondence ❑	7.10 manual / brochure ❑	7.11 advertising ❑	7.12 map, chart, table graph ❑
7.13 email ❑	7.14 Websites ❑	7.15 video ❑	7.16 CD Rom ❑
7.17 other text type(s) *(please specify):*			

8. And the book's text topics (heard, spoken, read, written)? (Please tick (✓) appropriate boxes ❑.)

8.1 accommodation ❑	8.2 physical environment ❑	8.3 social environment ❑	8.4 health environment ❑
8.5 daily routines ❑	8.6 food and drink ❑	8.7 shopping ❑	8.8 travel ❑
8.9 education, training ❑	8.10 world of work ❑	8.11 art ❑	8.12 literature ❑
8.13 music ❑	8.14 science, technology ❑	8.15 economics ❑	8.16 culture and customs ❑
8.17 current affairs ❑	8.18 other topics: *(please specify)*:		
8.19 Any inappropriate topics? *(please exemplify and explain)*:			

If the book has no recorded texts, please go to Item 10.

9. Authenticity of listening texts and tasks: (Please tick (✓) appropriate boxes ❑.)

9.1 Do the listening text(s) appear:	scripted? ❑	authentic? ❑	some of each? ❑
9.2 Do the recorded texts include redundancies such as:	repetition? ❑	rephrasing? ❑	hesitation? ❑
9.3 Please comment on the authenticity or realism of the listening tasks:			

10. Authenticity of reading texts and tasks: (Please *tick* (✓) the appropriate boxes ❑.)

10.1 Do the reading texts seem:	adapted or written for the book? ❑	authentic? ❑	some of each? ❑
10.2 Please comment on the authenticity or realism of the listening tasks:			

Most of the information you have been asked to provide so far has been relatively objective. Questions 11 and 12 here are very important as they request you to give your own more subjective evaluation of how the book treats the main language skill areas, and of the book as a whole.

11. Please give your <u>comments on the book's treatment of the four language skills</u>:

11.1 Listening	
11.2 Reading	
11.3 Writing	
11.4 Speaking	

12. Please now <u>evaluate the whole textbook</u>, preferably in terms of:

❑ type ❑ level ❑ contents ❑ pedagogical approach ❑ interest ❑ impact

13. Finally please evaluate the book's relationship with the international test for which you use it to prepare your students.

How does the book help your students to cope with the international test you are preparing them for?

That is the end of the questionnaire. Thank you very much for answering the questions.
RAH January 03

19 Crossing the bridge from the other side: the impact of society on testing

Cecilie Carlsen
Norsk språktest, University of Bergen

Abstract

This paper explores the relationship between testing and society from an untraditional angle, focusing on the effect **of** society **on** testing. Language testing in Norway, and more specifically the development and public reception of the national tests of English for Norwegian school children, is discussed as an illustration of this phenomenon.

Introduction

Test impact on society and individuals has been the subject of considerable research interest in the field of language testing during the last decades (Alderson and Wall 1993, 1996, Bailey 1996, Shohamy 2001, Wall and Horak 2006, 2007, 2008). The social consequences of test results are regarded as a central aspect of construct validity according to Messick's definition (Messick 1989), and the focus on test impact is claimed to distinguish modern language testing in the communicative paradigm from language testing before the 1970s (Bailey 1996).

When the relationship between testing and society is explored in our field, the focus is almost exclusively on the impact *of* testing *on* society. It does however seem a reasonable assumption that there is a two-way relationship between testing and society: not only do language tests affect society; language tests are also affected *by* society. The kind of society of which the tests are a part, affects test development, testing policy, the use of tests, as well as the public opinion about tests. In this paper the impact of society on testing will be discussed using Norway and the development of national tests for Norwegian school children as an example. I will start by describing Norway as an egalitarian society with strong socio-democratic traditions. Thereafter I will describe the Norwegian school system as a means to achieving equality and social mobility, and finally the role of testing within this system will be discussed.

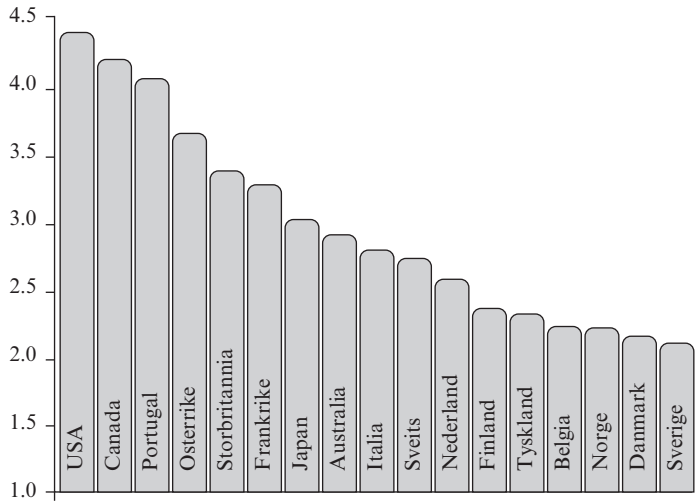
Norway – an egalitarian society with strong socio-democratic traditions

Norway and other Scandinavian countries are often described as examples of well-functioning welfare states, referred to as The Scandinavian Model. This model is usually associated with:

> [. . .] the State providing, financing and regulating welfare services for all citizens from cradle to grave. It is assumed to be the successful accomplishment of a strong and well organized labour movement of social democratic inclination; and it has been understood as a third or middle way between capitalism and communism (Abrahamson 1999: 32)

Important social tasks, such as education and health-care, as well as care for children and the elderly, are regarded as the state's responsibility, catered for at the public expense. There is low social inequality and poverty, a high level of employment including high female employment, gender equality, small differences in wages, and a tax policy based on the principle of condition, meaning that the more you earn, the higher the taxes on income. It has been a political aim to smooth out social differences, and to promote mobility between socio-economic classes: everybody should have the same rights and the same opportunities to succeed in life regardless of their background.

Figure 1 Differences in wages within the country, 19 OECD-countries (D9/D1 2006)



(Samlerapport – Kunnskapsdugnad for verdiskapning, 2007)

Economic equality

When compared with the other OECD-countries (Organisation for Economic Co-operation and Development) with regard to differences in salaries between the higher and lower earners, Sweden, Denmark and Norway are the three countries with the smallest differences, while USA, Canada and Portugal have the greatest differences (see Figure 1 above).

This policy of economic equality is strongly supported by the public opinion: investigations show that despite the existing small differences in wages, 70 % of the people want these differences in salaries to be reduced even further (Langeland and Stene 1999).

Investigations into people's perception of the society in which they live, show that most Norwegians believe that their society is one in which there is a fair distribution of economic resources. As Table 1 shows, 56% of Norwegians describe their society as one in which most people are to be found in the middle of the social pyramid. By comparison, only 12% of the French describe their society in the same way. Indeed, almost 50% of the French describe their society as a classical pyramid, having a small elite on top, many in the middle, but most on the bottom. This is also the view of most North Americans when describing their own society.

Table 1 People's perception of the society in which they live (Hjelbrekke and Korsnes 2006, p17 [own translation])

Kind of society	Norway	Sweden	Germany (West)	France	USA
A: A small elite on top, a few in the middle, very many on the bottom	3.1	10.3	10.6	12.4	16.2
B: Pyramid, small elite on top, many in the middle, most on the bottom	10.9	23.9	26.7	49.8	30.6
C: Pyramid, but few on the bottom	19.3	27.6	25.5	23.1	17.9
D: Most to be found in the middle	**56**	33	23.9	12.9	25.3
E: Many near the top, only few near the bottom	7.7	1.3	2.0	0.8	5.4
Don't know	.2	3.9	11.3	0.9	5.4
Total percentage	100	100	100	100	100
Total N =	(1,228)	(1,140)	(909)	(1,876)	(1,188)

Cultural equality

In addition to the economic equality described above, there is what we could call a principle of cultural equality in Norway; a preference for the average and a dislike for the extraordinary, i.e. people who are, or think they are,

better, wiser, fancier, more clever etc. than the rest. The McDonald's 'Crew person of the month'-kind of ranking is alien to the Norwegian way. People rather tend to be modest about their achievements and to play down their deeds. This way of thinking is captured in a concept from literature, the so-called 'Janteloven' (Jante Law), described by the Norwegian/Danish author Aksel Sandemose in 1933 in the novel *En flygtning krydser sit spor (A Fugitive Crosses his Tracks)*. Janteloven's main tenet is summed up in the phrase: 'Don't think you're anyone special or that you're better than us', and some of its rules are: 'Thou shalt not believe thou *art* something" and "Thou shalt not believe thou art greater than *we*'. The law describes a social phenomenon where people do not want to differ far from the norm. The Jante Law keeps people in their place. The phenomenon has its equivalence in the 'Tall Poppy Syndrome' in the UK, Canada, Australia and New Zealand, though I believe its influence has fallen far short of that of the Jante Law in Norway and Denmark, where the liking for average is to be found all the way back to our earliest literature, *Håvamål* (*The Words of the High One*), which reflects the wise words of the Norse god Odin, dating back to around year 800, which says in one of its verses (56):

> *It is best for man to be middle-wise,*
> *Not over cunning and clever:*
> *The learned man*
> *whose lore is deep*
> *is seldom happy at heart.*
> (Translated by P B Taylor and W H Auden 1969)

The Norwegian school system

Norway has a long tradition for a 'unitary school', which is a strong state-run, public, non-paying, anti-elitist school, where children from different social backgrounds meet in the same classroom. There are relatively few paying private schools in Norway, partly because of rather strict laws regulating this area. Less than 2% of Norwegian school children attend private schools (as opposed to roughly 15% of French, 12% of Danish, 10% of North-American and 7% of British school children). Because of the principle of equality, there are no marks in primary school (years 6–13), as it is an aim to differentiate between children as late as possible in their schooling.

Aim for school: result equality

It has been an important aim of the school system in Norway to wipe out any social differences, i.e. the school has been instrumental in promoting social mobility and socio-economic equality. This aim is explicitly stated in the introduction to the School Curriculum of 1997:

> Equality should be safeguarded between urban and rural areas, between social classes, genders, generations, between ethnic and linguistic groups and minorities, and across capability of functioning and across the range of abilities (Curriculum L97: 55, own translation).

In order to fully comprehend the meaning of this quote, it is necessary to take a closer look at the concept of *equality.* The concept may be defined in different ways and given somewhat different contents. The Norwegian professor of sociology and Minister of Education from 1990 to 1995 Gudmund Hernes (1974) distinguishes between four kinds of equality in relation to education*: Formal equality*, *resource equality*, *competence equality* and *result equality*. The first of these, *formal equality*, means that everybody should have the same chances to enter higher education, regardless of factors such as race, gender etc. There should be no differences *formally* as to who may access higher education. Yet, as long as parents have to cover the expenses, children from less advantaged homes will not benefit from this formal equality, and inequality and social differences between classes will be reproduced. The second kind, *resource equality*, means that the socio-economic situation of the parents should not influence their children's opportunities in life. Since not all parents are able to pay for their children's education the state should give financial support by providing a student loan and different kinds of scholarships. According to this principle of equality, everybody gets the same financial support – but again, Hernes claims, inequality is still present. Resource equality gives everybody a chance to participate in the same competition, but it does not compensate for differences in children's background. The third kind is *competence equality,* which means that more public finances are used on higher levels in education than on lower levels. The more effort a pupil makes, the more support he or she obtains. Clever pupils eventually receive more as they proceed in the educational system than the less clever students who drop out earlier, start to work and pay taxes which, in turn, finance the studies of the clever student. Again, inequality is the result. The fourth kind of equality, and the one which according to Hernes is the only kind that truly promotes social mobility, is *result equality*. The school system should not only give everybody the same chances, but compensate for differences in social background, by giving more to those with fewer socio-economic resources. This principle is reflected in Roemer (2000) who argues that:

> The ideal of the equal opportunity policy is to allocate educational resources to render it so that how well a person eventually does in the acquisition of the outcome in question reflects only his effort, not his circumstances (Roemer 2000:23).

To achieve this goal, more financial support is needed for pupils with a less advantageous socio-economic background, in order for them to reach the

same level of competence as the more advantaged children. As Hernes argues: 'Equality in results is ensured by inequality in the resources directed towards each pupil' (Hernes 1974: 249, own translation).

The role of testing within the unitary school system

It has been said of the Norwegian school system that it suits the average pupils, concentrates on the weak pupils, while the clever pupils are less stimulated to reach their full potential (Danbolt 2006, Andersen 2008).

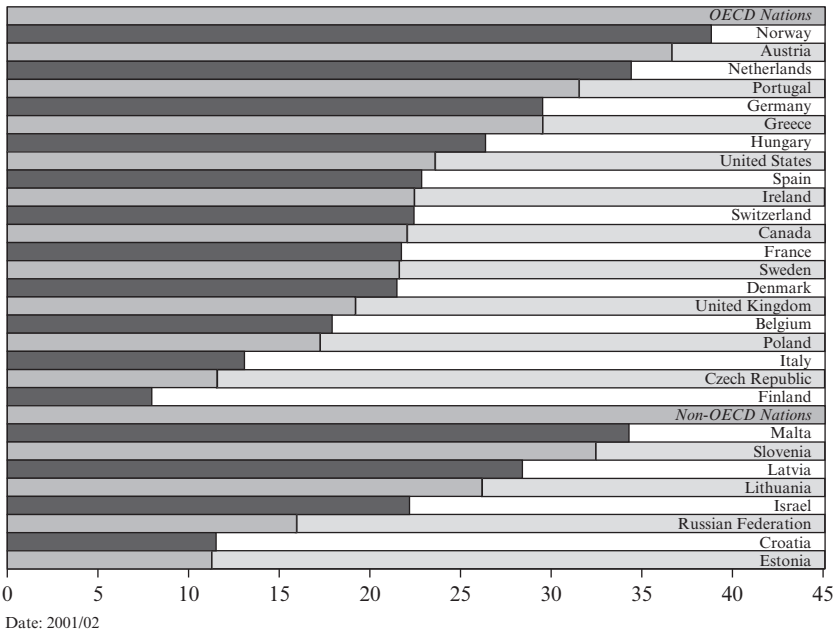
What is the role of testing within a school system where result equality and social mobility is the aim? Does testing have a role to play in a society where there is a strong preference for mediocrity, and where the Jante Law makes people loathe to stand out as excellent? Norway is of course a meritocratic society in the sense that scholarships, attractive jobs, and positions are distributed on the basis of qualifications and not on birthright. In society, then, testing has a role to play in achieving a fair distribution of privileges. But in the school system, and particularly in primary school, there has traditionally been very little testing compared to other European countries. In primary schools, the ideal of late differentiation between pupils has been dominant with no marks being given until secondary school (age 13). As a rule, testing in primary school has been limited to two kinds: firstly, control of whether pupils have done their homework, and secondly, standardised tests aiming at identifying pupils with reading or writing difficulties or other kinds of learning disabilities. The principle of result equality requires a means to detect which pupils are in need of more resources in order to obtain the same results as their peers. The standardised tests indicate where the extra resources are needed to give everyone an equal opportunity to perform well. Since the purpose of the standardised tests has been to identify pupils with learning difficulties, the tests are constructed to discriminate between the weak pupils and the others, but not between the average and the clever pupils, or between the clever and the very clever pupils. Consequently, primary school pupils, teachers and parents are not used to a kind of testing that challenges even the clever pupils.

The best school in the world – or not?

In Norway it has been the ambition of different political parties, socio-democrats and conservatives alike, to have the best school system in the world. Before 2000, the general opinion in Norway was that our school system was indeed a very good one. The results of the Programme for International Student Assessment 2000 (Lie, Kjaernsli, Roe and Turmo 2000) radically changed that view: PISA 2000 compared the reading skills of 15 year olds in 15 OECD countries. Norway performed averagely, just a little better than the OECD-mean (505 vs. 500 points), but not as well as Sweden (526 points), and far behind the

PISA-winner Finland (546 points). There were also positive findings, however: for instance Norwegian school children obtained high scores on social well-being at school, a finding supported by the UNICEF 2007 report: *An overview of child well-being in rich countries*. As Figure 2 shows, almost 40 % of Norwegian school children aged 11, 13 and 15 report that they 'like school a lot'. Actually, Norway is on the very top of the OECD-countries when it comes to well-being at school, and interestingly, Finland, the PISA-winner, is at the bottom. Only 7 % of the Finnish school children report that they 'like school a lot'.

Figure 2 Percentage of students age 11, 13 and 15 who report 'liking school a lot', UNICEF 2007 report



Clearly, then, the unitary school system has achieved some positive results. Norwegian school children may not read as well as the Finnish, but at least they really enjoy going to school. These positive results however received little attention, perhaps due to political changes which will be discussed further below.

Shift of government, shift of school policy

In 2001 a conservative government replaced the socio-democratic government. The conservatives wanted a shift in school policy, and they used

the negative PISA-results as a justification for change. The PISA-results showed clearly, they claimed, that the Norwegian unitary school was not good enough, and the laws regulating the area of private schools were liberalised making it easier to establish private schools (Friskoleloven 2003). The conservative government wanted competition between schools, and they needed a means of informing the public about school quality, giving parents the possibility to choose the best schools for their children, private or public. Through the introduction of a new curriculum the conservative government gave more freedom to teachers as to the content and methods of their teaching, but at the same time, introduced more control of learning outcomes. In other words there was less control of the *input* of teaching, but more control of learning *output* (Lieberg 2007). These two aspects, the need of informing the public about school quality on the one hand, and the need of controlling learning output, on the other, opened up for more testing in school, including in primary school.

National tests for Norwegian school children

The Minister of Education, Kristin Clemet, took the initiative to develop national tests for Norwegian school children in 2003. The proposal gained support in Stortinget (the Norwegian Parliament) (St.prp. nr. 1 Tillegg nr. 3, 2002–2003). The tests were to fulfil two different functions:

1. *Pedagogical function:* the tests should 'provide pupils, teachers and administrators with the necessary information to facilitate pedagogical development'
2. *Reporting function*: the tests should 'provide local and national authorities and the general public with information which can encourage dialogue and development of education standards' (Hasselgreen, Moe, Carlsen and Helness 2004).

It soon became obvious that there was strong tension between the two aims. For the test-developers the pedagogical aim was of paramount importance, which meant constructing tests with positive washback-effect, yielding useful and detailed feedback to teachers and pupils. This concern for the pedagogical consequences conflicted with the concern for the reporting function of the tests. Scores on school level were to be easily reportable on the internet, which meant that single scores were preferred to profiles and detailed feedback. In addition, in an early evaluation of the tests, high reliability estimates were given more importance than positive pedagogical consequences of the tests (Lie, Caspersen, and Björnsson 2004, Carlsen, Hasselgreen and Moe 2004).

Children were to be tested at four points in primary, secondary and upper secondary school, at grades 4, 7, 10 and 11, which means that they would be 9–10 years old the first time and 16–17 the last time they were tested. National

tests should be developed in four basic skills: reading in L1, writing in L1, English (reading and writing) as well as arithmetic. In this paper, I will concentrate on the development of the English tests.

National tests in English for Norwegian school children

The development of the national tests in English started out with strong pedagogical intentions within the team of test-constructors. (The article author was part of the team that developed the national tests in English from its start in 2003 until March 2005.) The national tests received a great deal of public attention, and they were felt to be high-stakes for teachers and school owners, so that the washback-effect on teaching could be assumed to be strong, according to Alderson and Wall's Washback Hypotheses (Alderson and Wall 1993). The concern for positive washback-effect on teaching and learning was considered of paramount importance to the group constructing the tests. The government initially wanted computerised tests that could be scored objectively, but out of concern for the washback-effect, the test constructors insisted on testing written production as well on all levels except in the 4th grade, since the teaching of English on this level is primarily focused on oral skills. For the other grades, writing tests consisting of three different tasks were developed. The assessment was based on a rating grid reflecting models of communicative competence, and based on the proficiency levels of the Common European Framework of Reference (CEFR). The rating grid was to be used as a guide for teachers when giving pupils and parents feedback about individual pupils' strengths and weaknesses. The politicians decided that it would be too expensive to pay a group of trained and experienced raters to rate the essays, so they wanted teachers themselves to rate their pupils. The test developers warned the politicians about the negative consequences of this procedure for the reliability of test scores, but this was not taken into consideration. However, the positive side effect of this procedure was that the all English teachers became familiarised with the CEFR and received training in assessing writing, an advantage much appreciated by the test developers. This was felt to be particularly important in primary schools where about 50 % of the teachers have no formal education in teaching English whatsoever (Drew 2004, Lagerstrøm 2007). Many teachers were therefore happy to learn more about assessing writing, and felt that the CEFR-based scale was a helpful tool in assessment as well as in informing pupils and parents. Another positive side-effect was that pupils in the 7th grade started to practise writing in English, something they had done very little of prior to the introduction of the tests.

In addition to the writing test, a computerised reading test was developed. This test was adaptive on test-level, meaning that pupils first got a series of items, and depending on their performance on these items, they were presented with a main test at a difficulty level adapted to their level of proficiency.

This meant that pupils in the same class would get somewhat different tests: the strong pupils would get a chance to perform their best, and the weaker pupils would not have to be confronted with a series of items too difficult for them. A lot of work was put into developing a large item-bank, piloting test items and standard-setting items to the CEFR levels.

Public reactions to the national tests

The introduction of national tests of writing, reading, English and arithmetic received a strong negative public reaction amongst teachers, parents and pupils, who organised action groups in order to put a stop to the tests. Pupils, particularly at upper secondary schools, boycotted the tests by staying home from school the days the tests were administered, and their boycott gained support from the parents' action groups.

It is perhaps not surprising that pupils are negative to testing in general, and to be subjected to a system of tests based on a political decision in particular. What is interesting about the negative reactions that arose is therefore not the negative response in itself, but the arguments that were used. The negative reactions were mainly due to the egalitarian ideal*:* there was fear that publishing the results on the internet would lead to the establishment of more private schools for those who could afford them, and an impoverishment of the state schools. In short, the negative reactions were to a large extent based on a fear that the national tests would contribute to increased differences between the rich and the poor (Berg 2004, Hølleland 2007).

In addition, teachers feared the extra work load involved, while parents and pupils feared that the national test would augment the load of stress on the part of the pupils. There were also some critical voices raising the question of negative washback-effect on teaching and learning, though the criticism showed little awareness that washback-effect can be positive as well as negative, depending on the test itself and on teaching prior to the test.

Many primary school teachers also claimed that the tests were far too difficult, even though the piloting of test items clearly showed that this was not the case. This is probably due to the fact that the other standardised tests known to teachers were the diagnostic tests referred to above, whose main function was to identify pupils with reading and writing difficulties. Primary school teachers, parents and pupils are used to tests where the average and the clever pupils get everything right and take fright at tests that are challenging for the stronger pupils.

Current testing situation in Norway

Globalisation has led to an increase in international pupil and student assessment programmes such as PISA. The results of such studies have made it clear

to Norwegian politicians that a system of quality control of learning output needs to be utilised on a regular basis. Even though we currently have a socio-democratic government with a school policy quite different from the one which introduced the national testing system, the National tests are still developed and administered. Politicians acknowledge the need of assessing children's basic skills of reading, writing, English and arithmetic, and realise that this cannot be done without tests which discriminate not only between the weak pupils and the others, but also between the average and the clever pupils.

Concluding remarks

The main concern of this paper has not been to criticise or to defend the national tests. Nor has it been my concern to defend or criticise those who oppose the introduction of the national tests. My main concern has been to demonstrate that testing in an egalitarian society like that of Norway is confronted with particular challenges. There is a large degree of opposition towards testing and grading in a society where equality is the aim. This is something of a paradox, testing being a crucial part of a democratic society. In a society where goods, positions and privileges are distributed by qualifications and not birthright, testing is an indispensable tool.

The traditional public opposition towards testing in Norway has acted as a brake on the professionalisation of the field of testing. It is still difficult to raise a professional debate regarding test quality, test ethics and a fair and reasonable use of test results in Norway. Rather it tends to stagnate in a discussion for or against testing altogether (Carlsen 2008). The conclusion of this paper, then, must be that international testing organisations such as ALTE (Association of Language Testers in Europe) and EALTA (European Association for Language Testing and Assessment) and its members have an important role to play in raising the consciousness about testing in Europe from parents and teachers to politicians.

References

Abrahamson, P (1999) The Scandinavian model of welfare, in Bouget, D and Palier, B (Eds) *Comparing Social Welfare Systems in Nordic Countries and France*, Paris: MIRE.

Alderson, J and Wall, D (1993) Does Washback Exist? in *Applied Linguistics* 1993 14 (2): 115–129, Oxford: Oxford University Press.

Alderson, J and Wall, D (1996) TOEFL preparation courses: a study of washback, in *Language Testing* 13 (3): 280–297.

Andersen, E (2008) På sitt nivå. Downloaded from <www.espen.com/norskblogg/archives/2008/05/p_sitt_niv.html> in August 2008.

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

Berg, K (2004) *En skole for økte forskjeller*. Downloaded from <www.intsos.no/?id=2191> in August 2008.
Carlsen, C (2008) Er testing skadelig? Kronikk in *Bergens Tidende* 13.02.2008.
Carlsen, C, Hasselgreen, A and Moe, E (2004) Kommentar fra Engelskprosjektet til rapporten "Nasjonale prøver på prøve". Downloaded from <www.regjeringen.no/nb/dep/kd/pressesenter/pressemeldinger/2004/Kommentar-fra-Engelskprosjektet-til-rapporten-Nasjonale-prover-pa-prove.html?id=103221> in August 2008.
Curriculum L97, *Læreplanverket for den 10-årige grunnskolen (L97)* (1997) Oslo: Kirke-, utdannings- og forskningsdepartementet.
Danbolt, S (2006) Å pugge seg ut av samfunnet. *Samtiden. Tidsskrift for politikk, litteratur og samfunnsspørsmål* 1- 2006, Oslo: H Aschehoug & Co.
Drew, I (2004) *Survey of English teaching in Norwegian primary schools*, Stavanger: Stavanger University College.
Friskoleloven (2003) Downloaded from <www.regjeringen.no/nb/dep/kd/dok/rundskriv/2003/Informasjon-om-den-nye-friskoleloven.html?id=109340> in August 2008.
Hasselgreen, A, Moe, E, Carlsen, C and Helness, H (2004) CATwalk to quality: The painstaking process of developing CEF-based computer-adaptive reading tests in Norway, presentation at EALTA Conference, Kranjska Gora, Slovenia, May 14–16, 2004.
Hernes, G (1974) Om ulikhetens reproduksjon:hvilken rolle spiller skolen?, in Mortensen, M (1974) (Ed.) *I forskningens lys* 147–170, Oslo: NAVF.
Hjelbrekke, J and Korsnes, O (2006) *Sosial mobilitet,* Oslo: Det Norske Samlaget.
Hølleland, H (2007) Nasjonale prøver og kvalitetsutvikling i skolen, in Tveit, S (Ed) *Elevvurdering i skolen. Grunnlag for kulturendring*. Oslo: Universitetsforlaget.
Lagerstrøm, B O (2007) *Kompetanse i grunnskolen. Hovedresultater 2005/2006.* Downloaded from <www.ssb.no/emner/04/02/20/rapp_200721/> in August 2008.
Langeland, O and Stene, R J (1999) *Holdninger til arbeid, lønn og fagbevegelse Resultater fra en spørreundersøkelse, Rapportserien i prosjektet Det 21. århundrets velferdssamfunn,* FAFO.
Lie, S, Caspersen, M and Björnsson, J (2004) *Nasjonale prøver på prøve*. Downloaded from <www.utdanningsdirektoratet.no/templates/udir/TM_Artikkel.aspx?id=1394> in August 2008.
Lie, S, Kjærnsli, M, Roe, A and Turmo, A (2001) Nasjonal hovedrapport PISA 2000: Godt rustet for framtida? Norske 15-åringers kompetanse i lesing og realfag i et internasjonalt perspektiv, in *Acta Didactica* No. 4. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.
Lieberg, S (2007) Den nye læreplantenkningen – hvilke utfordringer og muligheter gir den i opplæring av voksne*,* presentation at VOX-conference: Faglige dypdykk – læreplanen fra flere sider. VOX-konferansen om norskopplæring for voksne innvandrere. Oslo, 14.–15. mai 2007.
Messick, S (1989) Validity, in Linn, R *Educational Measurement,* New York: American Council on Education.
Roemer, J (2000) Equality of opportunity, in Arrow, K, Bowles, S and Durlauf, S (Eds) *Meritocracy and Economic Inequality*, Princeton: Princeton University Press.

Samlerapport – Kunnskapsdugnad for verdiskapning, 2007. Downloaded from <www.kunnskapsdugnad.no/rapporter.html> in August 2008.
Sandemose, A (1933) *En flygtning krydser sit spor*, Oslo: Tiden Norsk Forlag.
Shohamy, E (2001*) The Power of Tests. A Critical Perspective on the Uses of Language Tests,* London: Longman.
Stortingsproposisjon nr. 1 Tillegg nr. 3 (2002–2003) *Nasjonalt system for kvalitetsvurdering i grunnopplæringen.* Downloaded from <www.regjeringen.no/nb/dep/kd/dok/NOUer/2003/NOU-2003-16/19/5.html?id=370806> in August 2008.
Taylor, P and Auden, W (1969) *The elder Edda: a selection/translation from the Icelandic; introduction by Salus, P and Taylor, P,* London.
UNICEF 2007 report: *An overview of child well-being in rich countries*. Downloaded from <www.unicef-irc.org/publications/pdf/rc7_eng.pdf> in August 2008.
Wall, D and Horak, T (2006) The TOEFL Impact Study: Phase 1. The Baseline Study, *TOEFL Monograph* 34, Princeton, NJ: Educational Testing Service.
Wall, D and Horak, T (2007) Using Baseline Studies in the Investigation of Test Impact, *Assessment in Education* 14 (1): 99–116.
Wall, D and Horak, T (2008) *The TOEFL Impact Study: Phase 2. Coping with Change*. *TOEFL iBTResearch Series,* No. 05, Princeton, NJ: Educational Testing Service.

20 The educational and social impact of the CEFR in Europe and beyond: a preliminary overview

Brian North
Eurocentres Foundation
Chair, EAQUALS (European Association for Quality Language Services)

Abstract

The CEFR offers a common metalanguage to facilitate transparency and coherence in the provision of language learning in Europe and in the reporting of achievement in it. This paper reviews the effect that the CEFR is having on discussion of levels and comparison of language learning outcomes in Europe and beyond. The paper is organised in three sections. It starts by reminding the reader of the purpose and nature of the CEFR and points out the overall effect that it is having on professional networking. It then discusses policy impact, reporting from two recent surveys carried out by the Council of Europe's Language Policy Division and from the Language Policy Forum held in Strasbourg in 2007. Last but not least, it considers the practical impact of the CEFR, concluding with an assessment of its influence on examination reform and on the linking of language assessments in Europe.

Purpose and nature

The *Common European Framework of Reference for languages: learning, teaching, assessment* was published in final form in English and French in 2001 (Council of Europe 2001a, 2001b). There are two sides to the CEFR: on the one hand it is a compendium intended to offer a stimulus for reflection on and further development of current practice, and on the other hand it offers common reference points (levels and categories) to assist communication across educational sectors, national and linguistic boundaries. The CEFR makes clear that it is not intended as a harmonisation project:

> We have NOT set out to tell practitioners what to do or how to do it. We are raising questions not answering them. It is not the function of the CEF to lay down the objectives that users should pursue or the methods they should employ (Council of Europe 2001a:xi).

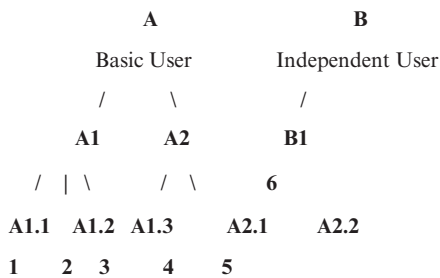
The approach taken is to provide a conceptual framework made up of:

- a taxonomic *descriptive scheme*, covering such issues as domains of language use, communicative language activities and strategies, plus the competences of the learner/user, based on the Council of Europe's work on objectives since the 1970s.
- a set of *common reference levels,* defining proficiency in as many of these categories as possible at six levels (A1, A2, B1, B2, C1, C2) in empirically developed scales of illustrative descriptors (North 2000a, North and Schneider 1998).

The idea is to offer a concertina-like reference tool for people to expand/contract and elaborate/summarise the levels and categories in order to adopt activities, competences and proficiency stepping-stones that are appropriate to the local context, yet can be related to the greater scheme of things and thus communicated more easily to colleagues in other educational institutions and to other stakeholders like learners, parents and employers.

Many people are familiar with the six 'Common Reference Levels' A1–C2. However, it is not so widely appreciated that, even though these six levels may have a certain self evidence as curriculum levels at least in a European context (Hargreaves 1992), the CEFR framework is a very flexible one. The levels are presented in a 'branching approach' with several examples included in Figure 1, inspired by the fact that the Finnish education system splits A1 into three levels.

Figure 1 A branching approach



This is not at all synonymous with the ALTE Framework, which is a notional grid of 24 cells (six levels by four skills − or 30 cells if one includes the language usage paper common to Cambridge examinations). ALTE provides examinations aimed at six specific levels across a wide range of

European languages. The CEFR, by contrast, provides a metalanguage to help people to communicate what the actual level of their different educational programmes and qualifications are.

In addition the CEFR offers a certain perspective that could be summarised by three points:

- **Transparency and coherence:** Define objectives and outcomes, create specifications and tell learners what they are. Relate the content of the programme or examination through 'Can Do' descriptors to the rest of the world, both in terms of the current and future real-life language needs of the learners concerned and in terms of the expectations of the educational sector to which they may be progressing.
- **Language for a social purpose:** See the learner as a social agent who needs to be able to perform certain actions in the language. Set realistic goals in certain activities in certain domains. In the third foreign language, for example, a receptive ability (follow TV or a presentation; read newspapers or professional papers) may be adequate.
- **Plurilingualism, cultural mediation:** Promote language learning as bridge-building – increasing international and intercultural understanding. Plurilingualism is primarily a question of perspective; limited, relevant competence in a range of relevant foreign languages is often more useful than all-round C2 competence in just one.

The 54 CEFR sub-scales for different communicative language activities and strategies and for different aspects of communicative language competence are relevant to all three points. They offer a means to specify the relevant terrain and to *profile* achievement in it.

The CEFR and professional networking

Networking in Europe: The CEFR has been translated into 36 languages. In many ways its face to the teaching profession is the European Language Portfolio (ELP), a reporting instrument intended to encourage reflection on and profiling of language learning in relation to CEFR-based descriptors. There are currently over 95 versions of the ELP that have been formally validated by the ELP Validation Committee. Whilst all share the same CEFR self-assessment grid, the descriptors on checklists tend to be simplified or further elaborated for specific contexts (university, lower secondary, young learners etc.). An extensive bank of such approved descriptors, collated by Günther Schneider and Peter Lenz, is available on the website www.coe.int/portfolio. There are at least two international versions for older learners, that of CERCLES, the European association of national university language centre associations, and that provided by EAQUALS/ALTE. The latter is available in an electronic version as a free download on www.eelp.org. The

Portfolio has become a considerable focus in teacher education and the subject of several European networking projects, organised under the EU Socrates programme and by the European Centre for Modern Languages (Little, Hodel, Kohonen, Meijer and Perclová 2007).

In addition, the CEFR has inspired an entirely new network of European language testers: EALTA (European Association for Language Testing and Assessment: www.ealta.org). More than 20 language testing agencies have gone as far as to conduct formal case studies trialling the procedures put forward in the pilot version of the Manual for relating language examinations to the CEFR (Council of Europe 2003), and have formed a loose network forged in four meetings between 2004 and 2007. EALTA itself, founded in 2004, had more than 250 attendees at its fourth conference in 2007 and the EALTA Best Practice Guidelines have been translated into 32 languages.

Why has the CEFR become popular in Europe? Goullier suggests that it is a case of lucky timing – or good planning:

> The realisation of the pressing need for educational instruments better geared to the variety of language skills required in Europe coincided chronologically with the CEFR providing the basics of a solution. In my view this coincidence accounts for the Framework's rapid success. Conversely, its success heightens awareness of the needs, insofar as it makes certain solutions technically possible (Goullier 2007:31).

Indeed the CEFR has been so successful in European ministries that a need has been felt to develop a similar common framework for the language of schooling, drawing on those aspects of the CEFR that seem relevant (Martyniuk 2007). Two prominent European language testers have offered their views as to why the CEFR has become so popular. Van Hest (2007) of the Dutch national language testing agency Cito suggests that the CEFR gives schools, teachers and students an international common framework to:

- discuss and promote language competence and language learning
- structure and plan the language learning process
- discuss progress and achievements in a transparent way.

Jones (2006:56), in discussing a framework for the 'Languages Ladder' needed for the English National 'Asset Languages' project, cites the conclusion of an English national report on problems with existing qualifications frameworks. Current qualifications are stated as being 'confusing and uninformative about the levels of competence they represented' (Nuffield Languages Programme 2002:8) and said not to support learning well, whereas a new framework should stress meaningful proficiency levels and provide a 'learning ladder' of bite-sized accessible learning targets. The report recommended the CEFR as a model for addressing both problems.

The CEFR has also been getting increasing attention outside Europe.

After a research and feasibility study (Vandergrift 2006), Canada has formally adopted the CEFR to provide a framework of common national reference points, whilst preserving the autonomy of diverse strategies at a provincial and local level. The core argument for the adoption of the CEFR in Canada has been summarised as follows (Macdonald and Vandergrift 2007):

- **Construct validity:** the CEFR level descriptors are based on a theory of communicative competence and empirically validated.
- **Face validity:** the level descriptors are congruent with teachers' perceptions and experiences with language learners (i.e. are not ivory tower applied linguistic constructs).
- **Contextual validity:** the branching approach to both levels and categories is able to accommodate the different needs and pedagogic cultures of the provinces and territories.

However, the authors did express concern that there was not enough differentiation at Basic levels to chart progress for beginner-level language learners. This is a common concern, but one beginning to be addressed to some extent by some current Portfolio reform projects in approaches similar to the Finnish one illustrated in Figure 1. In Switzerland, for example, a follow up project to the original CEFR/Portfolio descriptor research project (North and Schneider 1998), also at the University of Fribourg and using the same methodology, has produced a set of descriptors, tests and illustrative performance samples for the French and English performance of Swiss–German 14–16 year olds, published as Lingualevel (Lenz and Studer 2007, www.lingualevel.ch).

Several other countries have also started to adopt the CEFR. Taiwan, for example, has conducted a case study linking its national exams to the framework and required that international tests used in Taiwan report to the CEFR scale. Sixteen Spanish-speaking countries have recently collaborated to create SICELE (Sistema Internacional de Certificación del Español como Lengua Extranjera) that provides a common certificate related to the CEFR. The Foreign Languages Department of Osaka University has conducted a feasibility study on the adoption of the CEFR to give coherence to the curricula for the more than 20 languages that they teach. Both Osaka and the Japan Foundation have experimented with the use of the CEFR (which is available in Japanese) and the Portfolio for Japanese as a Foreign Language. In the United States, a network has been set up around an emulation of the ELP called 'Linguafolio' (Cummins, Davesne, Brinckwirth and Petillon 2007; www.doe.virginia.gov/linguafolio). From a separate US project, Bott Van Houten (2007) stated that the CEFR appeals in North America because it:

- makes language learning transparent
- motivates and enables the learner
- promotes reflective learning

- provides a new way of looking at culture
- recognises and values heritage languages
- documents individual performance
- facilitates articulation among language programmes
- provides a common criteria scale
- promotes language learning a lifelong endeavour.

However, the CEFR has had a long, slow germination over 20–30 years in the European modern languages world. The interest outside Europe at least raises the question as to whether the CEFR, 'Can Do' descriptors or an 'action-oriented approach' are appropriate and valid for pedagogic cultures and languages which were not considered at the time of its creation. This could be said to be an empirical question or an experiential question – experimentation with the Portfolio may indicate the degree of suitability.

Policy impact

It is probably true to say that nobody actually has a complete overview of either the policy or the practical impact of the CEFR at a local, national or international level. As regards the latter, CEFR levels are used for reporting in EU projects, such as EUROPASS and the upcoming survey of 15–18 year olds across Europe in the European Indicator of Language Competence. In an attempt to gauge the impact the CEFR is having, the Languages Policy Division of the Council of Europe recently carried out two surveys. The first, undertaken between March and June 2005, was a survey of institutions concerned with the teaching and learning of modern languages (Council of Europe 2006). The second, in 2006, was a formal survey of member states (Martyniuk and Noijons 2007).

2005 Survey of institutions: The 111 institutions who replied were situated in 39 states, including Mexico and Egypt. They classified themselves as follows: Higher education: 39; Central authority: 29; Teacher training/education: 36; Examination provider: 16; Language school/centre: 14; Adult education: 12; Other: Further education, publisher, primary or secondary school, cultural agency/centre: 28.

First the institutions were asked who used the CEFR and how useful they found it. It was said to be used mostly by teachers and teacher trainers, test writers and materials writers. There was a high consensus that the parts that were most widely known and most useful were the Common Reference Levels: understood as the global scale (Table 1), self-assessment grid (Table 2) and the illustrative descriptor scales. The survey also asked respondents to estimate on a 0–3 scale how useful the CEFR had been overall for their institution, and in particular how useful it had been for the following areas:

a. curriculum/syllabus development
b. pre-service teacher training
c. in-service teacher training
d. testing/assessment/certification
e. textbook writing/production of educational materials
f. communication with stakeholders (learners, parents, teachers, clients, etc.)
g. other contexts (please specify the context).

Average overall usefulness was 2.44 on the 0–3 scale with testing/assessment/certification (2.70) and curriculum/syllabus development (2.66) as the professional domains that most profited. The CEFR was stated to be most useful to examination providers (2.88). Next the survey asked respondents about problems with the use of the CEFR and elicited suggestions for further development. Among the comments were the following:

> *The CEFR is very promising in its philosophy and general idea, especially concerning multilingualism. Unfortunately it is very rarely used in the very sense of its own philosophy but is being misused as a simple testing instrument.*

> *Communication of and around the CEFR tends to focus on the global scales. Very little attention has been paid to the open, differentiated, descriptive nature of the sections of the CEFR which do not deal with levels and scaling. Furthermore, not enough attention has been paid to the differentiation within the scales (profiling). These factors have led to a generally prescriptive interpretation of the CEFR.*

The many comments made could perhaps be summarised as follows:

The CEFR is complex and relatively inaccessible, partly because it is a new approach:

- It is difficult to read straight through.
- It requires skills, study, and above all support: because it represents a fundamental shift in thinking; the next generation of teachers will find it easier.
- There is a need for drastic abbreviation and simplification if it is realistically to be used by today's teachers.
- People focus on the global levels. There is a general lack of awareness of the richness of the descriptive apparatus and of the implications for methodology.

The CEFR brings the risk of a prescriptive norm rather than the descriptive framework intended:

- There is a danger of harmonisation of pedagogic cultures, a 'globalisation' of approaches and methodologies, rather than the focus

on meeting needs in context and the respect for diversity that the CEFR seeks to promote.
- There is a tendency towards a simplistic interpretation of 'levels' without enough attention being paid to the differentiated use of the descriptive scales for profiling.
- There is a risk of automatic 'application' of the CEFR rather than its use as a dynamic tool for development.

Many suggestions were made, that can be structured into three groups.

Simplification: Differentiated use of the CEFR would be greatly helped if there was a simplified version/abstract or if practical guidelines were produced for teachers on aspects related to planning, teaching and assessment.

R & D networking: National workgroups or networks could be organised to create a cascade effect to spread a differentiated understanding of the CEFR and assist development. Such networks, and research programmes associated with them, could help to bridge the gap between academic researchers, faceless testers and school teachers working much as they did in the 1970s. In particular, projects could be organised to apply the action-oriented approach and to evaluate the use of the scales and descriptors in the classroom.

Further development: The 'CEFR Toolkit', (the various illustrative and supplementary materials available on www.coe.int/portfolio) could be developed in a number of ways. Variants on the core 'Can Dos' could be developed for children, working immigrants, etc.; in many cases such development could be supported by the Portfolio variants already developed, collated in the descriptor bank on www.coe.int/portfolio Materials illustrating specific 'Can Do' statements rather than samples illustrating the general level of proficiency could be created. Online training applications could be developed in all languages, including illustrative samples, exercises, self-assessment and proficiency test(s). At the time of writing www.CEFtrain.net is probably the only such site, and shows a modest example of the potential.

2006 Survey of member states: The second survey took place in 2006 as part of the preparation for the 2007 Intergovernmental Policy Forum discussed below. Thirty replies were received from the 46 states of the Council of Europe. The CEFR was said to be referred to at an official level in documents in the following areas:

- **Language policy:** strategy documents and action plans
- **Curricula:** for primary, secondary, higher education, bilingual education, minority languages

- **Assessment/Certification:** guidelines and requirements
- **Teacher training:** initial curricula and in service training
- **European Language Portfolio:** recommendation for its use
- **Textbooks:** guidelines for development
- **Migrants:** requirements for entry, residence, citizenship
- **Civil servants:** requirements for entry or for different grades.

Overall, the CEFR was considered to be having a major impact, having gained a wide acceptance: perceived as neutral, well known in institutions and quite well accepted by teachers. As a reference tool it was used in all sectors and its usefulness was widely acknowledged. It served as a development resource for policy documents, for curricula as well as, in some cases, for practical teaching or teacher training materials. There was stated to be a methodological impact in terms of initial and in service teacher training and in curriculum development. However, reflecting the comments made by the institutions the previous year, the CEFR was said not yet to be relevant to the teaching profession at a school level. The potential of the CEFR was not being realised because it is not reader-friendly. Very many language professionals are not even aware that it is a book. There is a need for mediation, simplification, explanation and training.

In particular, there was said to be a need for better clarification and exemplification in the form of comments on theoretical concepts, user-friendly summaries, example teaching/learning tasks for specific contexts, more dissemination in the form of international and national workshops, a forum for the exchange of good practice and, last but not least, the linking of national syllabuses and tests for different ages to the CEFR. With regard to the specific areas of curriculum development, assessment and certification, and teacher training, the main points made are summarised in the next three paragraphs.

Curriculum: The aspects of the CEFR felt to be especially useful were the learner-centred action-oriented approach, the scenarios offered for diversification of languages offered and the concepts of plurilingualism and pluriculturalism in a lifelong learning context. The main problems identified were the need to define additional sub-levels, a certain repetitiveness and lack of detail in some of the descriptors, a hesitance among teachers about the idea of 'partial competences' and – for some users – the lack of descriptors for mediation and translation skills. The main solutions to these problems suggested were:

- definition of appropriate sub-levels
- questionnaires to identify teachers' and learners' needs
- linking syllabuses, textbooks and examinations to the CEFR
- databases of model tasks for CEFR descriptors
- teacher seminars to raise competence and to exchange good practice.

Certification: The most useful aspects were stated to be the Common Reference Levels as reference points both for setting objectives and for assessing their achievement, the descriptor scales defining proficiency in different categories at different levels, and the framework and procedures provided (in the Manual for examination providers) to encourage the mutual recognition of language qualifications. The main problems identified were conservatism among teachers, with a need to involve parents and other stakeholders to overcome this, a certain lack of precision in some of the descriptors and, above all, the difficulty of linking tests to the CEFR. The main solutions to these problems suggested were:

- teacher training workshops on CEFR implementation
- the development of more illustrative descriptors and of sample test tasks
- the analysis of examinations by mixed teams of teachers and experts
- linking projects to relate syllabuses and examinations to the CEFR.

Teacher training: Use of the CEFR in teacher training is widespread, but the effects are difficult to summarise. The reference levels and the scales of illustrative descriptors were felt to be especially useful, including for defining the proficiency of teachers and trainee teachers themselves. There was, interestingly, little mention of the descriptive scheme as a way of conceptualising language learning and use, nor of the 'action-oriented approach'. The main problems identified were the lack of dissemination: teachers are simply not familiar with the CEFR, and the complexity – both of the actual document itself and of the theoretical concepts contained within it. The descriptive scheme and methodological approach proposed were described as difficult to access. Not many solutions were suggested. Those that were put forward included co-operation at an international level, and the reinforcement of teacher education through the demonstration of good practice in relation to the CEFR, possibly through assistance from publishers.

2007 Language Policy Forum: The issues raised in the 2005 and 2006 surveys as regards curriculum development, certification and teacher training were the focus of working groups at the Intergovernmental Language Policy Forum on 'The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities' held in Strasbourg, 6–8 February 2007 in honour of John Trim. As the report puts it:

> The objective was to offer the member states a forum for discussion and debate on a number of policy issues raised by the very speedy adoption of the CEFR in Europe and the increasingly widespread use of its scales of proficiency levels.

> This is because the clear success of the CEFR has significantly changed the context in which language teaching and assessment of language learning outcomes now take place in Europe. It was accordingly important to take stock of this new situation and to identify the key concepts behind the resulting dynamic (Council of Europe 2007:5).

Both in the preparatory meetings and in the actual Forum itself, there was a consensus that the CEFR essentially had two sides. Firstly the CEFR is an aid to reflection and reform – a compendium of approaches and elements to take into consideration, with chapters on aspects of teaching and learning, curriculum, assessment, with questions encouraging users to reflect on current practice. Secondly, it offers a set of common reference points – the levels, scales, descriptors – plus the samples illustrating them. There was a strong feeling that consideration of the former was being overshadowed by an often superficial focus on the latter. In this respect one must remember that the concept of a common European descriptor scale was first put forward as part of a presentation of the European Language Portfolio (North 1992) and was added to the concept of a Common Framework almost as an afterthought during the Rüschlikon Symposium. The main thrust of the discussion at both the Rüschlikon Symposium and in the CEFR Working Party (1992–96) concerned the role of the Framework as a reflective tool for curriculum reform and the elaboration of objectives based on the concepts of the 'action-oriented approach' and plurilingualism.

Discussion of policy direction at the Forum was thus far wider than just the issue of linking to the CEFR levels. Above all, there was a confirmation of the apparent trend towards considering problems and solutions through projects, meetings and networks at an international rather than national level. Goullier (2007) in his plenary talked of a 'shared European space' saying that '. . . the momentum generated by the CEFR in Europe is starting to bring about a new balance in terms of responsibilities, placing increased weight on horizontal relationships between member states and thus raising the issue of responsibility from a new perspective . . . A state's relationship with the Council of Europe can no longer be considered in isolation from what other states are doing' with regard to linking qualifications to CEFR levels or the development of CEFR-based content specifications (the 'Reference Levels') or in relation to curriculum reform, particularly with regard to the promotion of plurilingualism and intercultural competences.

Responsible collaboration: Much emphasis was given in presentations and discussion on the need to maintain the integrity of the validated descriptor scales and to preserve their value as an international standard and to take particular care when interpreting test results in CEFR levels. Users were exhorted to apply a quality assurance approach in so doing, to consider

setting up a national agency, and to above all link up with the various networks (e.g. EALTA, ALTE) able to assist in this process.

> If the Framework's reference levels are referred to in (*a state's*) education system or by organisations subordinate to it in order to assess learners' skills, each member state must be aware of its responsibility for ensuring that all the conditions are met for proper reference to be made to the Framework.
>
> Everything must be done to guarantee that the good practices identified at the international level for developing fair, transparent, valid, reliable examinations are adhered to. National or local authorities must take action to ensure that the levels of competence certified by their language examinations and the CEFR reference levels are linked in a transparent, reliable manner (Council of Europe 2007:14).

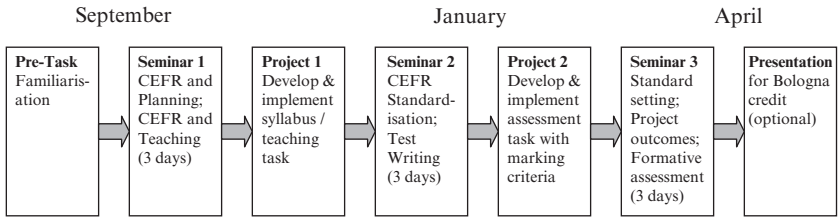
The success of the benchmarking seminars held to link examples of spoken performances to the CEFR scales (Lepage and North 2005, Norton and Lepage 2005), and of the portfolio projects that further developed CEFR descriptors for different contexts were both referred to, as well as the necessity to continue and extend such projects to further languages and educational sectors. A classified descriptor bank of CEFR-related descriptors from validated ELPs is on www.coe.int/portfolio. There was a wide consensus on the need for co-ordinated networking at national and international level with 'mutual shared responsibility': sharing reference points, expertise and tools whilst preserving the autonomy to design appropriate solutions for different contexts, objectives and needs.

***Accessibility for curriculum developers and teacher training*:** A second theme throughout the Forum was the issue of making the CEFR more accessible. There was much discussion of the need for practical guidance for curriculum developers, preferably with case studies showing the development of curricula taking account of the CEFR and/or of teaching materials illustrating the link with the framework. There was a need to develop and distribute training kits for teachers, materials illustrating the implications of the proficiency levels in different contexts, for different age groups and for different languages, and for documents making the CEFR more accessible for all teachers, while avoiding any reduction of its substance. This echoes the need stated in the two surveys for simplification, exemplification and guidance.

Simplification is, however, a complex matter. How does one simplify something which is intended to stimulate reflection by asking questions rather than providing answers? It is easier to do this well in a specific local context; it could be dangerous to do it at a global level. A multidimensional conceptual framework cannot be simplified without losing the multidimensionality that

facilitates the appropriate accommodation of diversity. One alternative to simplification is gradual training, for example in the type of modular 'sandwich course' recently organised in Switzerland for upper secondary teachers (Mayor and Péquignot 2008). The structure is illustrated in Figure 2.

Figure 2 Modular CEFR teacher training course



Differentiation: A third focus concerned the promotion of a more differentiated approach to teaching and learning modern languages. This primarily concerned three areas: 'partial competences', plurilingualism and the development of intercultural competences. The fact that most people are better at reception than at interaction or production in second or third foreign languages is still hardly reflected in official objectives, in teaching, or in language examinations. Secondly, modern languages tend to be taught in isolation from one another. The exploitation of transversal communicative and linguistic skills remains in its infancy; the recognition of and promotion of plurilingual profiles remains largely wishful thinking. In his short paper, John Trim put plurilingualism in its social rather than educational context.

> . . . The concepts of plurilingualism and pluriculturalism, largely developed by Daniel Coste, are of great value since they take a holistic view of linguistic and cultural competence.
>
> . . . This approach better meets the realities of globalisation than various forms of purism which regard each language and culture as a separate entity to be preserved and protected against the threat offered by alien forces.
>
> . . . Most users of the CEFR have applied it only to a single language, but its descriptive apparatus for communicative action and competences together with the 'can do' descriptors of levels of competence, are a good basis for a plurilingual approach to language across the curriculum, which awaits development (Trim 2007:49).

The paper now moves on to considering what practical impact the CEFR has had in the meantime, first considering curriculum, then language standards and finally relating examinations to common reference levels.

Practical impact

The CEFR's action-oriented descriptive scheme – *Activities and Strategies:* Reception/Interaction/Production/Mediation; *Competences*: Linguistic/ Pragmatic/Sociolinguistic/Intercultural – has at least the potential to replace Lado's (1961) four skills model – *Skills*: Listening/Speaking/Reading/ Writing; *Elements*: Grammar/Vocabulary/Pronunciation. However the impact of the descriptive scheme or other aspects of the CEFR on curriculum or teaching have as yet been very limited.

> To date (*the CEFR's*) impact on language testing far outweighs its impact on curriculum design and pedagogy. . .' (Little 2007:648) *and* 'On the whole the CEFR has no more occasioned a revolution in curriculum development than it has prompted the radical redesign of language tests (Little 2007:649).

The most that one can currently cite is a tendency to separate spoken interaction and spoken production in examinations (e.g. Goethe B2 and C1; the CIEP's DELF/DALF, Cambridge ESOL's reform of FCE (First Certificate) and CAE (Advanced) and the exams of the Spanish Escuelas Oficiales).

'Can Do's and curriculum: On the other hand, the wide dissemination of the CEFR 'Can Dos', mainly as checklists for particular levels in the ELP, has had some impact on curriculum design.

The actual quality of proficiency descriptors produced in Europe has also increased quite dramatically as people follow the concrete, 'salient feature' style used in the CEFR, as can be seen from the ELP descriptor bank on www.coe.int/portfolio Previously, vague statements for communicative activity of the type: '*Handles routine tasks with an adequate degree of competence*' had been common, and descriptors for qualitative aspects were noticeable for being negatively worded from levels B2 downwards – making them not much use as educational objectives (Trim 1978).

Since they are positively worded, the CEFR/ELP 'Can Dos' offer the potential to:

- relate objectives to real world needs and give a framework to action-oriented learning
- give a focus on specific micro-skills in listening and reading
- provide 'signposting' to learners, parents, sponsors
- relate assessment criteria for formal assessment procedures to CEFR descriptors.

Space does not permit a detailed exposition of all these points. Keddle (2004:45) has pointed out that at least EFL (English as a Foreign Language)

suffers from the comprehensibility of its 'mega coursebooks', which have caused the emphasis on authentic materials and microskills of the 1980s to be downplayed. 'Can Dos' for listening and reading can refocus teachers' attention on the acquisition of skills relevant to the level concerned, rather than just practice with comprehension questions.

More fundamentally, 'Can Dos' can be cross-referenced to objectives, syllabuses and materials in order to provide transparent 'signposting' in the form of a checklist of objectives for the term, a list of main communicative tasks and linguistic points in the current module, and a clear aim for the current lesson, for example written in an 'Aims Box' on the whiteboard. These approaches are common in EAQUALS – from a Greek primary school (aims boxes, checklists for teachers, report cards for parents) through language schools providing intensive courses in country and extensive courses at home (syllabus cross-referencing, checklists for teacher/ self-assessment) to a Turkish university (defining exit levels and detailed objectives, communication within faculty and with parents, continuous teacher and self-assessment). For example in Eurocentres every classroom has a standardised display of (a) the scale of levels, (b) the learning objectives for the CEFR level in question ('Our Aims') and (c) the objectives of the actual week's work (both communicative tasks and linguistic points). The weekly plan is introduced by the teacher on the Monday, and a review lesson at the end of the week combines a quiz on the main linguistic content with a small group discussion (oriented by photocopies of the weekly plan) of achievement of the week's objectives, and need for further class or individual work. Learners are thus treated as partners in the learning and teaching process.

To summarise, 'Can Do' descriptors give explicit signposting that help:

- teachers to explain syllabus choice – inviting comment and discussion
- learners to see why they are learning certain things
- learners and teachers to set priorities
- teachers to select appropriate communicative tasks
- learners and teachers to assess progress
- schools to report progress to parents.

CEFR levels as educational standards: The CEFR has led directly to the adoption of transparent standards for different educational sectors in many countries throughout Europe. In a project running up to the year 2000, Italy introduced portfolios and examinations targeted to CEFR levels developed by members of ALTE and CERCLU (the association of university language departments). In Finland, between 2001 and 2007 Sauli Takala and his team at the University of Jyväskylä related mainly teacher-centred assessments for five different educational sectors to the CEFR (Takala 2007). Between 2001 and 2005 Hungary implemented a thorough reform of language examinations,

linked to the CEFR (Alderson 2002). In Switzerland, which had launched the prototype Portfolio for ages 16+ in 2000, the CEFR inspired a national language concept which is currently taking concrete form in a project harmonising language standards for different educational sectors across the cantons. Lingualevel, the project that investigated the modern language competence of 14–16 year olds in the German-speaking cantons, and provided a separate Portfolio, DVDs of illustrative samples, assessment procedures and tests, has already been referred to (Lenz and Studer 2007; www.lingualevel.ch). In England the Asset Languages project adopted CEFR-related stepping stones and descriptors in order to create a 'languages ladder' for lifelong learning, (Jones 2006, Walker, Jones and Ashton 2007). The Netherlands has undertaken a curriculum reform and projects linking the main language examinations for different educational sectors to the CEFR (Kuijper and Noijons 2007). France adopted CEFR benchmarks for different educational sectors in 2006, together with a new focus on shared reflection about learning and an emphasis on practical, limited language use (*se debrouiller*), to be achieved though the use of workshop-style teaching (*ateliers*) and training in self-direction (*autonomie*). Germany set up a new national standards agency in 2006 that is developing a set of CEFR-related national language standards for different sectors across the federal states, similar to the Swiss harmonisation project.

Immigration: The CEFR has also been increasingly used as a reference point in defining the language level necessary for entry, residence and citizenship. The Threshold Level, now associated with B1, was originally conceived as a definition of the level an immigrant might aspire to, in order to be fully integrated into society. Certain countries (e.g. Finland) have taken this as a reference point for citizenship. Both the Netherlands and France, however, have recently given official recognition to a Level A1.1 – halfway to Level A1, which in the CEFR text itself was referred to as 'Tourist'. In France, the DILF (Diplôme Initiale en Langue Française) is an examination course developed mainly for immigrant women who may be becoming literate through French. The idea is that after this initial achievement they will move onto the ladder of DELF modules, starting with DELF A1. The Netherlands has adopted Level A1.1, assessed through an automated telephone test, as the minimum level for entry purposes. There is some controversy over the use of CEFR levels to define standards for immigration in this way. One should bear in mind that the CEFR did not give rise to the practice of setting language hurdles in this way and that it at least offers a way to translate vague notions like '*be fluent*' or '*be able to express themselves*' into concrete proposals that can be discussed in a meaningful way, compared to policy in other countries and, if necessary, criticised for lack of realism. The Council of Europe's Language Policy Division has run a project from 2003 to 2008 to define a *Language Policy for the Integration of Adult Migrants.* A

code of practice has been developed with help from EAQUALS and a survey of national policies has been carried out in 2008 with help from ALTE. The project aims towards providing a bank of relevant CEFR/ELP descriptors plus appropriate DVDs illustrating the CEFR levels in this context.

Linking examinations: Initially, the response of language professionals to the CEFR was to estimate the level of language qualifications to CEFR levels on the basis of individual expert judgement. In the context of saying that sensible methods of linking existed, and that even a private language school organisation like Eurocentres had been applying them in simple ways for years, North (2000b:556) complained that '. . . it is surprising the extent to which most of the other (*than ALTE*) national and international initiatives appear to rely solely upon committee decisions and self-regulation with little apparent qualitative analysis let alone quantitative analysis'. In his report on the Hungarian examination reform project, Alderson (2002:14) added '. . . and . . . seem prepared to content themselves with unprofessional and indefensible off-the-cuff assertions about levels of difficulty, of test tasks that have not even been empirically validated, let alone subjected to rigorous standard-setting exercises . . .'. With the formal publication of the CEFR in 2001, the issue became more urgent. Responding to requests for guidance, and to an offer from the Finnish education authorities, the Language Policy Division therefore organised a seminar in Helsinki (Council of Europe 2002), which defined the outline approach then taken in the pilot version of the Manual on linking examinations to the CEFR (Council of Europe 2003).

The pilot manual proposed procedures in three sets:

- **Specification** of the format, content and procedures of the examination, plus of coverage in relation to the CEFR categories and levels.
- **Standardisation** of the interpretation of the levels through training with illustrative samples followed by standard-setting/benchmarking of local samples to the framework levels.
- **Empirical validation** of the results through an independent corroboration of the linking claimed on the basis of specification and standardisation.

This scheme was adopted (a) because these categories are a good way of grouping linking methodologies found in the literature, (b) because they reflect the classic three stages of quality management (design, implementation, evaluation) and (c) because such broad concepts could thus be applied equally to formal, high-stakes assessment situations (examinations) and to lower-stakes school and teacher assessments.

The draft Manual was piloted in some 20 case studies throughout Europe (Martyniuk, forthcoming). Feedback indicated that the structure (specification, standardisation, validation) worked well, but that standard-setting required more detailed treatment. Comments suggested that the Manual

was a good way to critically review and evaluate the content and the statistical characteristics of an exam – and many stressed that the process is as important as the outcome. Walker et al (2007:4) suggested that although the Manual was stated to be 'not a guide to how to construct valid tests' this was an understatement, since its central concern – how to make valid criterion-referenced interpretations of test performance (to the CEFR) – is the central problem of test validity. They added that the Manual has educative potential for quality assurance, particularly for smaller or less well-resourced operations, as alignment to the CEFR required integration into every stage of the design and administration cycle.

The combination of the CEFR and the Manual can certainly be claimed to have had an impact, having helped institutions to:

- design an exam systematically so that results relate to an international standard: (e.g. CIEP: DELF/DALF and TCF; City & Guilds: Pitman exams)
- define performance criteria systematically (e.g. CIEP, City & Guilds, German KMK, Eurocentres, Escuelas Oficiales, Sabanci University)
- reflect on the role and composition of panels of experts (e.g. City & Guilds, Trinity College)
- move beyond 1970s-style standard-setting by guesswork to creative and critical corroboration of cut-scores (e.g. Cito; TestDaf; CIEP; City & Guilds; ECL Hungary).

As a result, the position in 2008 is a very different one to the situation in 2002. There is now a wide acceptance of the Manual's philosophy of 'building a validity argument'. The feedback of experience through use of the Manual in the 20+ case studies has raised interest and commitment across the continent. EALTA has grown into an organisation as significant as ALTE and we are now moving into a period of more specialised workshops, like the EALTA/Cito Seminar on standard-setting held in Athens in May 2008 and the Council of Europe/CIEP Séminaire interlangues/ Cross-linguistic Benchmarking Seminar held in Sèvres in June 2008 (Breton, Lepage and North 2008). The result is a substantial increase in expertise in Europe in the development of examinations and in linking them to an external standard.

Conclusion

As regards the question of the overall impact of the CEFR, to quote Ho Chi Min's supposed response when asked about the impact of the French Revolution: it is too early to say. A common framework is a social construct, a constructed consensus; the CEFR descriptors are a scaling of current shared perceptions of proficiency – based on an objective scaling of inter-subjective

judgements. In the technical literature on linking assessments, standards-oriented assessment (Gipps 1994) based on a common framework of standards is called social moderation (Carey 1996, Linn 1993, Mislevy 1992). The CEFR, its descriptors and the increasing toolkit of analysis tools and illustrative samples have been accepted because they accord with the perceptions of European language teachers, testers and language policy professionals. The process of fixing precisely what the CEFR levels mean cannot be separated from the networking, workshops with the illustrative samples and collaborative empirical studies that are now happening. It is a social process. The CEFR is not the revealed truth, and it is certainly not the last word. The current descriptors are illustrative. The purpose of the CEFR is to encourage reflection not to shut down debate.

References

Alderson, J C (2002) *Relating national examinations to the Common European Framework: The Hungarian experience*, unpublished paper.

Bott Van Houten, J (2007) *Language learning policies in the United States*, paper given at the intergovernmental Language Policy Forum 'The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities,' Strasbourg, 6–8 February 2007.

Breton, G, Lepage, S and North, B (2008) Cross-language benchmarking seminar to calibrate examples of spoken production in English, French, German, Italian and Spanish with regard to the six levels of the *Common European Framework of Reference for Languages (CEFR)*, CIEP, Sèvres, 23–25 June 2008, Report, CIEP/Council of Europe, Language Policy Division, available at: http://www.coe.int/T/DG4/Portfolio/documents/Report%20cross%20language%20benchmarking%20seminar-%20FINAL%2012%20Jan%2009%20(2).doc.

Carey, P A (1996) *A review of psychometric and consequential issues related to performance assessment*, *TOEFL Monograph Series* MS -3, Princeton, NJ: Educational Testing Service. September 1996.

Council of Europe (1992) *Transparency and coherence in language learning in Europe: Objectives, assessment and certification, symposium held in Rüschlikon, 10–16 November 1991, (Report edited by North, Brian)*, Strasbourg: Council for Cultural Co-operation.

Council of Europe (2001a) *Common European framework of reference for languages: learning, teaching, assessment*, Cambridge: Cambridge University Press.

Council of Europe (2001b) *Cadre européen commun de référence pour les langues: Apprendre, enseigner, évaluer*, Paris: Didier.

Council of Europe (2002) *Seminar on relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEF), Helsinki, 30 June 30–2 July 2002: Report*, DG IV / EDU / LANG (2002) 15, Strasbourg: Council of Europe.

Council of Europe (2003) *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)* DGIV/EDU/LANG (2003) 5, Strasbourg: Council of Europe.

Council of Europe (2006) *Surveying the use of the Common European Framework*

of Reference for Languages (CEFR) Draft synthesis of results, December 2005, Strasbourg: Council of Europe. DGIV/EDU/LANG (2006) 2.
Council of Europe (2007) *The Common European Framework of Reference for Language (CEFR) and the development of language policies: challenges and responsibilities,* Intergovernmental Language Policy Forum, Strasbourg, 6–8 February 2007, Report, available at http://www.coe.int/t/dg4/linguistic/Forum07_webdocs_EN.asp#TopOfPage.
Cummins, P, Davesne, C, Brinckwirth, A and Petillon, C (2007) *Future directions for LinguaFolio and the European Language Portfolio*, paper presented at the annual meeting of the American Council on the Teaching of Foreign Languages, Henry B Gonzalez Convention Center, San Antonio, TX, Nov 12, 2007.
Gipps, C (1994) *Beyond testing*, London: Falmer Press.
Goullier, F (2007) Impact of the Common European Framework of Reference for Languages and the Council of Europe's work on the new European educational area, in Council of Europe (2007) *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities.* Intergovernmental Language Policy Forum, Strasbourg, 6–8 February 2007, Report: 29–37.
Hargreaves, P (1992) Round table discussion on the European Language Portfolio, in Council of Europe 1992: 150–158.
Jones, N (2006) The impact of the CEFR on language testing in Europe, *Proceedings of the Japan-Europe International Symposium 2006 'A new direction in foreign language education: The potential of the Common European Framework of Reference for Languages*, Osaka University of Foreign Studies, Osaka, 5th March 2006. Osaka: OUFS Committee for Educational Improvement, 52–59.
Keddle, J S (2004) The CEF and the secondary school syllabus in Morrow (Ed.) *Insights from the Common European Framework*, Oxford: Oxford University Press, 43–54.
Kuijper, H and Noijons, J (2007) *Mapping the Dutch foreign-language examinations onto the CEFR*, paper given at the intergovernmental Language Policy Forum 'The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities', Strasbourg, 6–8 February 2007.
Lado, R (1961) *Language Testing. The Construction and use of foreign language tests: A teacher's book*, London: Longman.
Lenz, P and Studer, T (2007) *Lingualevel. Instrumente zur Evaluation von Fremdsprachenkompetenzen 5. bis 9. Schuljahr*, Bern: Schulverlag. www.lingualevel.ch
Lepage, S and North, B (2005) *DVD de productions orales illustrant pour le français les niveaux du Cadre européen commun de référence pour les langages du Conseil de l'Europe*, DVD, CIEP / Eurocentres: Paris, published with Council of Europe (2001b).
Linn, R L (1993) Linking results of distinct assessments, *Applied Measurement in Education* 6 (1), 83–102.
Little, D (2007) The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy, *Modern Language Journal* 91 (4): 645–655.
Little, D, Hodel, H, Kohonen, V, Meijer, D and Perclová, R (2007) *Preparing teachers to use the European Language Portfolio: arguments, materials and*

resources, Languages for Social Cohesion, Strasbourg: European Centre for Modern Languages/Council of Europe.

Macdonald, J and Vandergrift, L (2007) *The CEFR in Canada*, paper given at the intergovernmental Language Policy Forum 'The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities,' Strasbourg, 6–8 February 2007.

Martyniuk, W (Ed.) (forthcoming) Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's draft manual, *Studies in Language Testing*, Cambridge: Cambridge University Press.

Martyniuk, W (Ed.) (2007) *Towards a Common European Framework of Reference for Languages of School Education? Proceedings of a conference*, Krakow. Describing and Testing Language Proficiency 12, Universitas.

Martyniuk, W and Noijons, J (2007) *Executive summary of results of a survey on the use of the CEFR at national level in the Council of Europe Member States*, paper given at the intergovernmental Language Policy Forum 'The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities', Strasbourg, 6–8 February 2007.

Mayor, B and Péquignot, V (2008) Der GER in Zentrum der Ausbildung: Bericht über einen Einführungskurs, *Babylonia* 2/08, Contone, Fondazione Lingue et Culture: 53–54.

Mislevy, R J (1992) *Linking Educational Assessments. Concepts, issues, methods and prospects*, Princeton NJ: ETS.

North, B (1992) European Language Portfolio: Some options for a working approach to design scales for proficiency, in Council of Europe 1992, 158–174.

North, B (2000a) *The Development of a common framework scale of language proficiency*, New York: Peter Lang.

North, B (2000b) Linking language assessments: an example in a low-stakes context, *System* 28, 555–577.

North, B and Lepage, S (2005) Seminar to calibrate examples of spoken performances in line with the scales of the *Common European Framework of Reference for Languages*, CIEP, Sèvres, 2–4 December 2004, Report, Strasbourg, Language Policy Division, Council of Europe, DGIV/EDU/LANG (2005) 1, available at: http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/illustrationse.html.

North, B and Schneider, G (1998) Scaling descriptors for language proficiency scales, *Language Testing* 15 (2), 217–262.

Nuffield Languages Programme (2002) *A learning ladder for languages: Possibilities, risks and benefits.* Retreived from: http://languages.nuffieldfoundation.org/filelibrary.pdf/learning_ladder.pdf. Cited by Jones 2006.

Takala, S (2007) *Using the CEFR for assessing language competences in Finland*, paper given at the intergovernmental Language Policy Forum 'The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities', Strasbourg, 6–8 February 2007.

Trim, J L M (1978) *Some possible lines of development of an overall structure for a European unit / credit scheme for foreign language learning by adults*, Strasbourg: Council of Europe.

Trim, J L M (2007) The CEFR in relation to the policy aim of the Council of Europe, in Council of Europe (2007) *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities. Intergovernmental Language Policy Forum, Strasbourg, 6–8 February 2007, Report*: 48–49.

Vandergrift, L (2006) *New Canadian perspectives: Proposal for a Common Framework of Reference for Languages for Canada*, Canadian Heritage.

Van Hest, E (2007) *The CEFR in the Netherlands*, paper given at the intergovernmental Language Policy Forum 'The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities,' Strasbourg, 6–8 February 2007.

Walker, T, Jones, N and Ashton, K (2007) *Asset Languages: A case study of piloting the CEFR Manual*, paper given at the seminar for a joint reflection on the use of the preliminary pilot version of the Manual for 'Relating language examinations to the CEFR' 2004–2007: Insights from case studies, pilots and other projects, Cambridge, 6–7 December 2007.

Notes on the volume contributors

Oksana Afitska is a Research Assistant in the Graduate School of Education, University of Bristol, UK. Her research interests and expertise lie in areas of second language acquisition and language assessment, in particular formative classroom assessment. She is currently engaged in two research projects, SPINE (*Student Performance in National Examination: the dynamics of language*) and Research and Evaluation of the Diploma Qualification Project (QCA).

Rachel L Brooks serves as Applied Linguist for the US Federal Bureau of Investigation (FBI), working in the Language Testing and Assessment Unit where she oversees speaking testing and contributes to the development of language tests. Her research interests include translation performance, native speaker performance on speaking tests, and Forensic Linguistic methods to determine cheating.

Paula Buttery is a Senior Research Associate in the Research Centre for English and Applied Linguistics at Cambridge University, UK. She is also an Associate Researcher in Cambridge University's Computer Laboratory and a Research Associate in the Department of Linguistics, University of California Davis. Her research interests lie in the fields of natural language engineering, corpus linguistics, cognitive modeling and computational neurolinguistics. She is a member of the corpus group within the English Profile Programme and has worked on providing tools for searching corpora as well as providing analysis of extracted data.

Cecilie Carlsen holds a Master's degree in second language acquisition and a PhD in language testing. Since 2005 she has been working with test development and validation at Norsk språktest, at the University of Bergen in Norway, where she is engaged as a Postdoctoral Research Fellow.

Micheline Chalhoub-Deville is Professor of Educational Research Methodology at the University of North Carolina, Greensboro, USA, where she teaches courses in second language assessment, measurement and research methodology. She has published widely in the field, including many journal articles and two edited volumes: *Issues in Computer Adaptive Testing of Reading Proficiency* (1999) and *Inference and Generalizability in Applied Linguistics: Multiple Perspectives* (2006).

Jamie Dunlea has worked in English language teaching and testing in Japan since 1992. Since joining Japan's Society for Testing English Proficiency (STEP) Foundation in 2002, he has worked as a part of the team which produces the EIKEN English language tests and has been involved in a number of research projects, including the EIKEN Can-do List.

Mark Garner is co-ordinator of postgraduate coursework in linguistics and language teaching at Aberdeen University, UK. He has held similar positions elsewhere in Britain, Australia and Indonesia and has published a number of books and articles on theory and method in applied linguistics, operational communication, and research methods.

Roger Hawkey has many years of experience in English language teaching, teacher education, course design, and assessment projects in Africa, Asia and Europe. He is now a consultant on testing with Cambridge ESOL and a Visiting Professor with the University of Bedfordshire, UK. He has published widely in applied linguistics, language teaching and assessment.

John A Hawkins is Professor of English and Applied Linguistics at Cambridge University, UK, and Professor of Linguistics at the University of California, Davis in the USA. He has published extensively on English language typology, grammatical theory and psycholinguistics. He has been leading one of the research strands of the English Profile Programme, examining the Cambridge Learner Corpus.

Dayong Huang is currently a full-time PhD student in the University of Aberdeen, Scotland. He is also an associate professor in the Civil Aviation Flight University of China. His research interests are in language testing and aviation English. He has published a number of articles in the two fields.

Stergiani Kostopoulou is currently completing her PhD in Trinity College Dublin's English Language Support Programme and is an instructor of Applied Linguistics. Her research offers a CEFR-based curriculum for English language support for immigrant students in Irish post-primary schools using a corpus-based methodology.

Geraldine Ludbrook is a researcher at the University of Venice, Italy, where she teaches both undergraduate and postgraduate courses. She is also a teacher trainer at the Veneto teacher education institute (SSIS). She is currently a PhD student in Applied Linguistics at the Department of Linguistics and English Language, Lancaster University, UK.

Denise Lussier is a specialist in educational psychology and in measurement and evaluation, and a Professor at McGill University, Montreal, Canada. She co-ordinates a research project on *Cultural representations, ethnic identity and intercultural competence*. She has co-coordinated two projects for the European Centre for Modern Languages: 1) *Cultural Mediation and Language Teaching, 2*) *Guidelines for the assessment of intercultural communicative competence.*

Beth Mackey is a manager for the US Department of Defense. She holds an MA in TESOL and an MA in Russian and East European Studies. Her research interests are in the field of listening, specifically transcription as a language testing tool. She is co-chair of the Interagency Language Roundtable (ILR) Testing Committee.

Brian North is Head of Academic Development at Eurocentres, the Swiss-based language school foundation. He is co-author of the CEFR, the prototype European Language Portfolio, and the CEFR examination manual as well as the developer of the CEFR descriptor scales and Chair of EAQUALS, the European school accreditation scheme.

Szilvia Papp is a Validation Officer at the University of Cambridge ESOL Examinations, UK. She has worked on language, migration and citizenship issues since 2006 with academics and language testing professionals, presenting at international conferences. Currently she is conducting a study on the impact of language testing for migration purposes.

James Purpura is Associate Professor of Linguistics and Education in the TESOL and Applied Linguistics Programs at Teachers College, Columbia University in New York, USA. Besides directing these programmes, he teaches courses in language assessment and research design. He has published extensively in language assessment journals and is author of two books: *Learner Strategy Use and Performance on Language Tests: An SEM approach* (1999) and *Assessing Grammar* (2004). Jim was President of the International Language Testing Association (ILTA) in 2007 and 2008.

Pauline Rea-Dickins is Professor of Applied Linguistics in Education at the Graduate School of Education, University of Bristol, where she is also Director of Research. She has published widely in areas of language testing and assessment, particularly classroom-based assessment and language programme evaluation. She has worked extensively in Sub-Saharan Africa and is director of a major research project *Student Performance in National Examination: the dynamics of language*, a collaboration between Bristol and the State University of Zanzibar.

Wayne Rimmer is a doctoral candidate at the University of Reading, UK. His thesis explores the density and range of subordination in the International Corpus of Learner English (ICLE). He has taught English as a foreign language in Moldova, Russia, Thailand, Germany and the UK, and is currently co-authoring a pedagogic grammar for Cambridge University Press.

Philida Schellekens is an independent consultant and researcher in the field of language learning and teaching. She is an associate OFSTED inspector and teaches English part-time. Recent publications include *The Oxford ESOL Handbook* (2007) OUP, *Language in Construction* (2004) CITB/CILT, and *English Language as a Barrier to Employment, Education and Training* (2001) DfES.

Philip Shawcross holds degrees from the universities of Manchester, UK, and Toulouse, France. He has been an Aviation ESP teacher since 1974 involved in course development, auditing, CBT/WBT design, teacher training and consulting. He is the Director of English for Aircraft (UK) and training curriculum director of Aviation English Services (NZ). He is the Co-founder and President of the International Civil Aviation English Association (ICAEA).

Lynda Taylor is a Consultant to University of Cambridge ESOL Examinations and formerly Assistant Director of the Research and Validation Group there. She has extensive experience of the theoretical and practical issues involved in L2 testing and assessment, and has provided expert assistance for test development projects worldwide. She regularly teaches, writes and presents on language testing matters and has edited and written chapters for several of the volumes in the *Studies in Language Testing* series.

Margaret van Naerssen holds a PhD in Applied Linguistics and Language Acquisition from the University of Southern California, USA. She is currently at Immaculata University in Pennsylvania, working on international assignments with the US Department of State. Since 1997 she has undertaken expert work in forensic linguistics (at federal and state level) relating to legal cases of murder, rape, drugs, money laundering, robbery, and fraud, most involving non-native speakers of English.

Cyril J Weir holds the Powdrill Chair in English Language Acquisition at the University of Bedfordshire, UK, and is Guest Professor at Shanghai Jiao Tong University, PRC. He has taught short courses and carried out consultancies in language testing, evaluation and curriculum renewal in over 50 countries worldwide. He has published many books on language testing, including *Language Testing and Validation: an evidence-based approach* (2005), *Examining Writing* (2007) and *Examining Reading* (2009). He is also

joint Series Editor of *Studies in Language Testing*. Current interests include academic literacy and test validation.

Guoxing Yu is a Lecturer in Applied Linguistics at the Graduate School of Education, University of Bristol, UK. His main research interests are educational assessment, including language testing, evaluation and monitoring of school effectiveness in relation to language and literacy development, and measurement of learning power. Currently he is working on the SPINE project, and is also directing an IELTS funded research project, *The Cognitive Processes of Taking IELTS Academic Writing Task One*.

Presentations at the ALTE Conference Cambridge, 2008

Kate Green and Lid King
Dept for Children, Schools and Families, UK
Developing an assessment system in support of national policy

Gisella Langé
Italian Ministry of Education
Which competences and qualifications for the CLIL teacher?

Cecilie Carlsen
University of Bergen, Norway
Crossing the bridge from the other side: the impact of society on testing policy

Sacha DeVelle
University of Cambridge ESOL Examinations, UK
Language proficiency and testing for migration purposes: what are the practical implications?

Griet Ramaut and Machteld Verhelst
Katholieke Universiteit Leuven, Belgium
Testing academic language proficiency at the start of primary education: a functional language test

Robert Raabe
Universidad Complutense de Madrid, Spain
Developing an Erasmus exam

Miriam Sester Retorta
Federal University of Technology – Paraná, Brazil
Washback effects of public university entrance examination on high school environment in Paraná, Brazil

Eeva Tuokko
Finnish National Board of Education, Finland
What levels do 9th graders reach in terms of CEFR and Finnish National Core Curriculum?

Marita Härmälä
University of Jyväskylä, Finland
Evaluation de la compétence langagière dans les qualifications basées sur la compétence

Lorenzo Rocca
Università per Stranieri di Perugia, Italy
Il progetto per l'elaborazione del CELI Impatto I

Antony Kunnan
California State University, USA
Citizenship in the 21st century and its implications for citizenship assessment

Eli Moe
University of Bergen, Norway
Linking tests to the CEFR: is the present Manual a sufficient linking tool?

Konrad Schröder, University of Augsburg, Germany and Claudia Harsch, Institute for Educational Progress (IQB), Germany
Bridging the gap: how to introduce rating techniques into the FL classroom

Anne Gallagher
National University of Ireland
A new beginning: an examination of the washback effect of tests in Irish for adult learners

Sharon Jordan, University of Cambridge ESOL Examinations, UK, Béatrix Sampsonis, Alliance Française, France and Stefanie Steiner, Goethe-Institut, Germany
Introducing new BULATS Online, prefaced with a general introduction to BULATS in English, French and German

Jamie Dunlea
Society for Testing English Proficiency (STEP), Japan
The EIKEN Can-do List: improving feedback for an English proficiency test in Japan

Alec Johns
Cambridge Assessment, UK
The art of testing: variations on the predictable

Karen Lund
University of Aarhus, Denmark
The CEFR and the Danish C1 exam – a decontextualised linguistic approach

Geraldine Ludbrook
University of Venice, Italy
Certifying teachers' language proficiency: developing a performance test for CLIL teachers in Italy

Thomas Christiansen
Università del Salento, Italy
Fluency and pronunciation in the assessment of grammatical accuracy in spoken production. An empirical study

Silvia Irimiea and Livia Bradea, Babeş-Bolyai University, Romania
The assessment of English for tourism purposes. Reconciling national standards with EU standards.

Pauline Rea-Dickins, University of Bristol, UK, Matthew Poehner, Pennsylvania State University, USA, Constant Leung, King's College London, UK, Lynda Taylor, University of Cambridge ESOL Examinations, UK and Elana Shohamy, Tel Aviv University, Israel
Breaking the mould: evaluating validity from a situated language assessment perspective

Philida Schellekens
Independent Consultant, UK
The Skills for Life Strategy in England and Wales: washback on language learning for migrants and refugees

Martha Galvis, British Council Colombia and Ken McIntyre, ABC English Institute, Colombia
Aligning Colombian national tests with the Common European Framework of Reference for Languages

Stergiani Kostopoulou
Trinity College Dublin, Ireland
Self-assessment throughout the European Language Portfolio: raising the educational standards

Ardeshir Geranpayeh
University of Cambridge ESOL Examinations, UK
The application of SEM to enhance the revision of high stakes assessment: the case of FCE in English examination revision

Gerardo Valazza
Instituto Cultural Anglo Uruguayo, Uruguay
The impact of Cambridge ESOL TKT on language teachers, language schools and society in Uruguay

Marianne Mavel
Centre International d'études Pédagogiques (CIEP), France
Evaluation de la formation des correcteurs du DELF

Roger Nunn
Petroleum Institute, UAE
Designing rating scales for task-based and project-based learning

JoAnne Neff van Aertselaer
Universidad Complutense de Madrid, Spain
Language assessment for the teaching and learning of academic writing

Rosalba Rizzo
Centro Linguistico d'Ateneo Messinese (CLAM), Italy
Il Portfolio Europeo dell Lingue: cambio di prospettive nella valutazione del docente

Liu Yang
National Educational Examinations Authority, China
Test method effect on writing performance of Chinese students

Margo Gottlieb, World-Class Instructional Design and Assessment, USA and Neil Jones, University of Cambridge ESOL Examinations, UK
Application of language assessment to teaching and learning across continents: validation of 'Can Do' descriptors

Michaela Perlmann-Balme
Goethe-Institut, Germany
Einstufungstest-system für Integrationskurse – The development of a test for placement into language courses for migrants in Germany

Margaret Allan
Scottish Qualifications Authority
The SQA Framework of National Qualifications in ESOL: supporting social inclusion through language assessment

Waldemar Martyniuk
Jagiellonian University Kraków, Poland
State Certificates in Polish as a Foreign Language: impact and stakeholders

David Newbold
University of Venice, Italy
Co-certification: a new direction for external assessment?

Sandy Forster
Edmonton Public Schools, Canada
Bridging curriculum and international expectations: linking Canadian provincial language curricula to the CEFR

Radmila Bodrič
University of Novi Sad, Serbia
Standardised Foreign Language Testing in Serbia: a necessity, not an option

Maria Davou
Lancaster University, UK
Can-do: what and how can they do after all?

Immaculada C Báez Montero and M Rosa Pérez Rodríguez, Universidade de Vigo, Spain
La competencia lingüística en las pruebas de acceso a la Universidad Espaňola

Mohd Sallehhudin Abd Aziz
Universiti Kebangsaan Malaysia
Malaysian University English Test: its impact on the stakeholders at micro and macro levels

Suhair Al-Alami
Al Ghurair University, UAE
Diagnostic assessment in ELF contexts

Maryanne Hall and Kate Bellamy, National Information Recognition Centre (NARIC), UK
EUROPASS How can students fulfil their potential in Europe?

Neil Jones
University of Cambridge ESOL Examinations, UK
Building a framework for lifelong learning – Asset Languages

Margarete Schlatter, Federal University of Rio Grande do Sul, Brazil, Matilde VR Scaramucci, State University of Campinas, Brazil and Silvia Prati, University of Buenos Aires, Argentina
Celpe-Bras and CELU proficiency exams as political acts in Brazil and Argentina

Mónica Pereña, Generalitat de Catalunya, Spain, Pablo Sanz and Laura Riera, Universitat Autònoma de Barcelona, Catalunya, Spain
An online placement test that detects levels of general foreign language proficiency (simtest)

Lorna Carson
Trinity College Dublin, Ireland
Assessment in an institution-wide modern languages programme: using the Common European Framework of Reference

Peter Mickan
University of Adelaide, Australia
Making sense of tests and the semiotics of assessment

Codruța Goșa Romania and Luminița Frențiu, Universitatea de Vest Timișoara
Is it worth taking it? – attitudes to two high stakes English language exams in Romania

Rama Mathew
Delhi University, India
What can university students do with language?

Emyr Davies
Welsh Joint Education Committee
Developing A1 and A2 tests for a less widely used language: Welsh

Ulrike Arras
TestDaf-Institut, Germany
Die Erforschung von Beurteilungsstrategien und ihre Bewusstmachung durch Schulungsmassnahmen als Voraussetzungen für die Testvalidität

Lai Auyeung Winnie Yu Win
University of Hong Kong
Chinese Language teachers' assessment literacies and its impact on using learning portfolio to enhance learning

David Graddol
The English Company (UK) Ltd, UK
Future demographics of English learners

Alexis Colombia and Gerriet Janssen, Lopez Universidad de los Andes, Colombia
CEFR-based testing in non-European countries: test development and validation

Ilse Born-Lechleitner
Johannes Kepler Universität, Linz, Austria
Testing for student orientation – developing an online orientation test for first term university students

Hilary Maxwell-Hyslop
Chartered Institute of Linguists, UK
Maintaining professional standards and meeting the needs of stakeholders – assessing interpreting skills

Jackie Greatorex
Cambridge Assessment, UK
Exploring the role of human judgement in examination marking: findings from some empirical studies

Shalini Roppe
University of Leuven, Belgium
The Common European Framework of Reference: is alignment with functional performance based language examinations possible?

Stuart Wachowicz
Edmonton Public Schools, Canada
Building social and educational demand for second language education through internationally recognised language proficiency

Betty Lanteigne
American University of Sharjah, UAE
Societal effects of English language testing and policy in the Arabian Gulf

Dominique Casanova, Chambre de Commerce et d'Industrie de Paris (DRI/E-CCIP), France
Suivi et évaluation d'impact d'un test international de français langue étrangère: entre synergies et tensions

Meilutė Ramonienė and Loreta Vilkienė, Vilnius University, Lithuania
Lithuanian L2 speakers: winners or losers in a changed socio-political situation?

Marcella Binchi
Trinity College London, Italy
Monitoraggio sulle motivazioni e gli effetti della certificazione Trinity nella scuola italiana

Kieran O'Loughlin
University of Melbourne, Australia
Language standards: how are they viewed by the Australian ELT sector?

Tony Lee, Hong Kong Institute of Education and Wall Street Institute International, Hong Kong and Simon Buckland, Wall Street International, USA
The Wall Street Institute Can-Do study – a rasch calibrated curriculum alignment with ALTE levels

Roger Hawkey
University of Cambridge ESOL Examinations, UK
A study of CPE exam washback on textbooks in the context of Cambridge ESOL exam validation

Yan Jin
National College English Testing Committee, China
Criterion-reference testing in the reform of ELT at the tertiary level in China

Fiona Barker
University of Cambridge ESOL Examinations, UK
Using corpora for language assessment: trends and prospects

Rachel Lunde Brooks
Federal Bureau of Investigation, USA
In your own words: using authorship attribution to identify cheating on translation tests

Victoria Crisp and Nadežda Novakovic, Cambridge Assessment, UK
The effects of features of examination questions on the performance of students with dyslexia

Dina Tsagari
Hellenic Open University, Greece
Revisiting the concept of language washback: Results from an empirical study

Joëlle Crowie, Università degli Studi di Bologna, Italy and Liz McIlvanney, University of Bologna, Italy
Reading tasks for the assessment of language proficiency: an Italian case study

Pilar Concheiro Coello Iceland and Erlendina Kristjansson, University of Reykjavik, Iceland
Business Spanish, Legal English and Business English in Iceland: assessment methods and the business sector

José Pascoal
University of Lisbon, Portugal
Portuguese language exams: CIPE, DEPLE, DIPLE, DAPLE, DUPLE: background and present development of a tool for language policy

David Little
Trinity College Dublin, Ireland
The European Language Portfolio and adult migrants: an alternative approach to teaching and assessment

Simon Lebus
Cambridge Assessment, UK
Intelligent regulation: risk and trust

Michaela Perlmann-Balme, Goethe-Institut, Germany, Gilles Breton, Centre International d'études Pédagogiques (CIEP), France and Giuliana Bolli, Università per Stranieri di Perugia, Italy
"All different – All equal?" Towards cross language benchmarking using samples of oral production in French, German and Italian

Ana-Maria Pascu
Babeş-Bolyai University, Romania
Metacognitive strategies for BEC Higher

Szilvia Papp
University of Cambridge ESOL Examinations
The language requirements of the UK test for citizenship: Is it a valid test?

Roger Bowers
Trinity College London, UK
"Define plagiarism in your own words": linguistic larceny and its place in learning and assessment

Margaret van Naerssen
Immaculata University, USA
Going from language proficiency to linguistic evidence in court cases

Linda Mesh
University of Siena, Italy
The impact of the assessment of Business English: a case study of Italian teachers and learners

Anne Tangy
Association pour la Formation Professionnelle des Adultes, France
Lisibilité at reconnaissance des compétences linguistiques dans la qualification professionnelle

Jo Lewkowicz, Warsaw University, Poland and Elżbieta Zawadowska-Kittel, Higher School of Linguistics in Warsaw, Poland
English for all: the impact of the new school-leaving examination of English in Poland

Denise Lussier
McGill University, Canada
Empirical study and validation of a conceptual framework for the development of intercultural communication competence

Anthony Green
University of Bedfordshire, UK
Functional Progression in the Common European Framework: towards a language specification for higher level learners

Thomas Eckes and Gabriele Kecker, TestDaf-Institut, Germany
Putting the Manual to the test: the TestDaF-CEFR linking project

Corien Stevens
Cito, Netherlands
The use of portfolios in immigration courses in The Netherlands

Barbara Spinelli and Francesca Parizzi, Università per Stranieri di Perugia, Italy
Bridging language assessment and language teaching: the Reference Level Descriptions for Italian (A1–B2 Level)

Gill Elliot, Nat Johnson and Sylvia Green, Cambridge Assessment, UK
"Aspects of Writing": using an atomistic approach to evaluate qualities of features of writing

Bart Deygers and Dominique Neyts, Ghent University, Belgium
A multi-layered approach to reviewing tests of languages for specific purposes (LSP)

Margarete Schlatter, Federal University of Rio Grande do Sul, Brazil and Matilde VR Scaramucci, State University of Campinas, Brazil
Certificate of Proficiency in Portuguese as a Foreign Language: Impact in Brazil and overseas

Marian Amengual, University of the Balearic Islands, Spain and Honesto Herrera, Universidad Complutense de Madrid, Spain
The influence of the Spanish University Entrance Examination on the teaching of English

Gerardo Prieto and Juan Miguel Prieto, Universidad de Salamanca, Spain
Procedimiento para anclar un examen del DELE con el Marco de Referencia

Marylin Kies
Università degli Studi di Siena, Italy
Motivating toward certification

Marguerite Kuzma
European Commission
L'Indicateur Européen des Compétences Linguistiques et son contexte

Evelina Galaczi and Mick Ashton, University of Cambridge ESOL Examinations, UK
Assessment for teaching: Cambridge ESOL's CLIL exam

Bregje van Oel
University of Amsterdam, Netherlands
How to test integrated language skills

Sylvie Lepage, Centre International d'études Pédagogiques (CIEP), France, Bernd Tesch, Institute for Educational Progress (IQB), Germany and Roselyne Marty, Centre International d'études Pédagogiques (CIEP), France
Standardisation of the evaluation of written production by the IQB and the CIEP

Evdokia Karavas and Xenia Delieza, University of Athens, Greece
On site observation of KPG oral examiners: Implications for oral examiner training and evaluation

Hyeonjeong Jeong, Tohoku University, Japan, Shuken Shiozaki, Society for Testing English Proficiency (STEP), Japan and Ryuta Kawashima, Tohoku University, Japan
On site observation of KPG oral examiners: Implications for oral examiner training and evaluation

Tomoki Matsudaira
Society for Testing English Proficiency (STEP), Japan
Investigating Japanese learners' study habits and actual use of English

Rachida Guelzim
British Council Rabat, Morocco
The washback effect of assessment in Moroccan schools

Julia Todorinova
St. Kliment Ohridski University of Sofia, Bulgaria
A developmental perspective on standardized language testing at the Department for Language Teaching (Sofia University)

Cristina Pérez-Guillot, Miguel Angel Candel Mora and Asunción Jaime Pastor, Universidad Politécnica de Valencia, Spain
El uso de la lengua maternal en la elaboración de pruebas de nivel

Dayong Huang
University of Aberdeen, UK
Cheating in language testing: a case study of the CET in China

Viacheslav Godik, The Leading Center of Russian Language Testing for Foreigners, Russia and Tatyana Perova, The Leading Center of Russian Language Testing for Foreigners, Russian Federation
System of Testing Russian as a Foreign Language in the Russian Federation: recent achievements

Cyril Weir
University of Bedfordshire, UK
Developing equivalent forms in reading examinations: construct, construct and construct

Jamie L Schissel
University of Pennsylvania, USA
The impact of practicality constraints on the administration of test accommodations for English language learners

Dianne Wall
Lancaster University, UK
The TOEFL Impact Study: The effects of the new TOEFL on classroom teaching

Stuart Shaw
University of Cambridge International Examinations, UK
Proposing a new English IGCSE syllabus

Carlos Soler Montes
Instituto Cervantes, Calgary, Canada
Variación lingüística y evaluación: El caso de los Diplomas de Español como Lengua Extranjera (DELE)

Alistair Fortune
British Council, Bosnia-Herzegovina
Standards in STANAG 6001 Testing – A Peacekeeping perspective

Shigeru Ozaki
Takushoku University, Japan
The impact of entrance examinations on foreign language education policy: high school teachers' perceptions

Nükte Durham
Middle East Technical University, Turkey
Theory in action: a Turkish experience of the socio-cognitive model of speaking assessment

Nkechi Christopher
University of Ibadan, Nigeria
Social and educational impact of language assessment in Nigeria

Alma Oritz
Universidad Nacional Autónoma de México, Mexico
El impacto del MRCE en la certificación de ingles en la Universidad Nacional Autónoma México

Martine Derivry
Université Pierre et Marie Curie, France
Le biais culturel dans l'évaluation en langues: L'anglais dans le cadre d'un examen français

Philip Shawcross
International Civil Aviation English Association (ICAEA)
Social, safety and economic impacts of global language testing in aviation

Catriona Scott
Newcastle University, UK
Stakeholders' views of impact from high-stakes testing in the UK EAL/ESL primary context

Roger Dunne and Adriana Abad Florescano, Universidad Veracruzana, Mexico
The impact of CEFR-based foreign language assessment in Mexico: the EXAVER project

Michael Corrigan
ALTE Validation Unit
ALTE Validation Unit and diffusion of good testing practice

Tony Lee
Hong Kong Institute of Education and Wall Street Institute International, Hong Kong
Developing an EFL reading ability scale using Rasch measurement models v test design and calibration issues

Beth Mackey, US Department of Defense, USA and Rachel Lunde Brooks, Federal Bureau of Investigation, USA
Testing the LCTLs in the US Government: when is a bad test better than no test?

José Ramón Delgado Castillo
International School of Havana, Cuba
Accountability in EFL assessment: an experience at the International School of Havana

César Luis Díez Plaza
Instituto Cervantes, Belgrade, Serbia
El comcepto de "nativo" vs "no nativo" en las pruebas de certificación

Jessica Wu
Language Training & Testing Center, Taiwan
Views of some Taiwanese students and teachers on English language testing and assessment

Wayne Rimmer
Reading University, UK
Operationalizing linguistic creativity

Maria Eugenia Obrist
Academia Argüella, Argentina
The impact of language assessment in Argentinian education

Pauline Rea-Dickins and Guoxing Yu, University of Bristol, UK
Student performance in national examinations: the dynamics of language in school achievement

Johanna Motteram and Peter Mickan, University of Adelaide, Australia
Issues in test preparation and language development

Karen Ashton
University of Cambridge ESOL Examinations, UK
Studying impact in a new assessment framework

Danilo Rini, Università per Stranieri di Perugia, Italy and Michael Corrigan, ALTE Validation Unit
Implementation of an Italian item bank for large-scale test construction

Lucy Chambers
University of Cambridge ESOL Examinations, UK
Computer-based and paper-based writing assessment: A comparative text analysis

Junko Majima and Antonio Smith, Osaka University, Japan
Educational Impact of the CEFR on a Japanese national university

Rubina Gasparyae, American University of Armenia (AUA) and Karine Muradyan, The Castle Learning Center, Armenia
Bridging the past and the future in the field of testing and assessment in Armenia

Khaled El Ebyary
Newcastle University, UK
Addressing the complexity of washback

Shao Chin
Language Training & Testing Center, Taiwan
Contaminating reading with writing? On the use of summary writing tasks in EAP assessment

Qian Dai
Communication University of China
Washback of open-ended composition – a study of Senior High School English writing instruction in Beijing

Trisevgeni Liontou
University of Athens, Greece
Reading comprehension module of the Greek state exam – the test takers' perspective