ResearchNotes

Contents

The URL for reading/downloading single articles or issues of *Research Notes* is:
www.CambridgeESOL.org/rs_notes

The URL for subscribing to *Research Notes* is:
www.CambridgeESOL.org/rs_notes/inform.cfm

Editorial Notes

Welcome to issue 17 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

The theme of this issue is language testing in a European context, focussing on the work of the Association of Language Testers in Europe (ALTE) of which Cambridge ESOL is a founding member. This issue will describe the history and activities of ALTE, concentrating on some of the many projects being undertaken by ALTE members which impact on many thousands of language learners in Europe and further afield.

The introductory article by Barbara Stevens outlines the constitution and work of ALTE, concentrating on how its members work together to develop ways of describing and comparing their language examinations and to establish common levels of proficiency to aid the transferability and portability of language qualifications. She reports on a number of research projects including the ALTE Framework Project and work on the ALTE Code of Practice and Quality Assurance.

Peter Hardcastle reports on a project to investigate three European language tests of Reading at Threshold level (B1) in terms of describing the equivalence of the tests themselves and the constructs they test by comparing their task types and formats. This is an ambitious project and an important aspect of the work of ALTE members. Peter also describes the challenges faced by users of the Common European Framework of Reference when applying such a tool for comparing tests across languages and having to balance its theoretical and practical aspects in the process.

In the following article Simon Fenn describes how an ALTE and EAQUALS electronic language learning portfolio is being developed with the help of Cambridge ESOL. This project will enable language learners to record their aims and successes and is one way in which ALTE/EAQUALS members can motivate language learners to record and celebrate their achievements. Staying with how technology is influencing language learning and testing, Stuart Shaw considers how writing assessment can be automated and reviews four computerised systems for marking essays and short-answer questions. This area is a key focus for Cambridge ESOL at present and could radically alter the rating process in future.

We also include the Editors' notes from the latest Studies in Language Testing volume, number 18, which contains the key papers presented at the first ALTE Conference held in Barcelona in 2001. This is followed by conference reports focussing on reading, writing and English for Academic Purposes attended by staff, together with a report of the Young Learners Symposium recently hosted by Cambridge ESOL.

We end this issue with a range of Other News items and a call for proposals for the second ALTE conference taking place in Berlin in May 2005. We look forward to this stimulating event and hope to see you there.

A Common Solution to a Common European Challenge: The work of ALTE

BARBARA STEVENS, EUROPEAN PROJECTS OFFICER, RESEARCH AND VALIDATION GROUP

Introduction

With the breaking down of international barriers between European states and increasing opportunities for the members of the workforce to move from one country to another, the need for the transferability and portability of language qualifications is unmistakable. First and foremost, employers and employees alike need to know what language qualifications gained in various countries mean – what the holder of a given certificate can actually be expected to do – and how to make meaningful comparisons between qualifications gained in different states of a now enlarged European Union. Employers need to know which language qualification it is realistic to demand when advertising a post, and employees have an interest in being able to rate their own present level of expertise and future training needs. Consequently, the members of ALTE, the Association of Language Testers in Europe, have been working together to devise a means of describing and comparing their language examinations to establish common levels of proficiency in order to promote transnational portability and recognition of language certification in Europe.

What is ALTE?

ALTE is an association of providers of European foreign language examinations, which includes many of the major international bodies in the field of language testing (see list of members on page 5). It was founded in 1990 by the University of Salamanca and Cambridge ESOL (then UCLES EFL), and in 2004 now has 26 full members and 2 associate members, with membership growing continually. ALTE members come from a wide variety of organisations, including universities, ministries, state and private language training institutions. Some represent languages spoken widely in the world, others the so-called less widely spoken languages. As such there is a wide variety of interests within the association, and members are able to learn from each other.

ALTE is constituted as an EEIG – European Economic Interest Grouping – under European Law and this has determined certain aspects of the way it works. ALTE was set up, like other EEIGs, to create a legal entity based on EU law to facilitate and encourage cross-border co-operation. When setting up the association, it was found to be difficult for some potential members to enter into associations under foreign rules or jurisdictions and so the use of procedures under European law was thought appropriate. An EEIG, in fact, may be set up in any one of the EU Member States, and operate in any part of the EU but it can also enter into arrangements with organisations outside the EU, although these organisations cannot themselves become full members of an EEIG. ALTE's rules on membership eligibility derive from this restriction: full membership of the association is only open to organisations from member countries of the European Union and European Economic Area. However, ALTE also admits Associate Members as well as Observers (from outside the EU), the main difference being that the Associate Members do not have voting rights within the formal arrangements of the EEIG. The use of the word "economic" in the title of the grouping can be misleading. The ALTE EEIG is not a commercial organisation and does not have commercial gain as the group's objective. In fact, in the regulations for establishing an EEIG it is made clear that an EEIG itself cannot be formed with the object of making a profit. In EEIGs, like ALTE, where non-profit making organisations are involved, their contribution is generally seen in terms of "the services and skills they can provide".

ALTE can thus rightly be regarded as an "interest grouping" which carries out particular activities for the benefit of the members and a wider public audience. In the case of ALTE, the member institutions are almost all publicly-funded organisations such as ministry departments, cultural institutions, universities, and research institutes, or other not-for-profit organisations with charitable status.

Furthermore, ALTE became an International Non Governmental Organisation with participatory status in the Council of Europe in 2003. Participatory status is granted to organisations which are particularly representative in their field of competence at a broader European level and are capable of contributing to and participating actively in Council of Europe deliberations and activities.

ALTE Research Projects

Since the foundation of the organisation the members of ALTE have been working together to develop a framework for describing and comparing their language examinations and for establishing common levels of proficiency. This goal is being pursued by means of an ongoing 'ALTE Framework Project'. The first stage of this project was achieved in 1991 with the production of the first version of an ALTE Handbook which contained descriptions of the examinations offered by all the ALTE members. All of the examinations included were described using the same format, and details given of recognition of educational or workplace qualifications, the relevant importance attached to different skills, the testing focus of each paper, and the number and types of questions used.

Can Do Statements

A central platform of this long-term research project is constituted by the development of the ALTE Can Do Statements. The aim of the Can Do project being 'to develop and validate a set of performance related scales, describing what learners can actually do in the foreign language' (Jones and Hirtzel 2001:244.) The Can Do system, comprising approximately 400 statements

organised into three general areas: *Social and Tourist, Work*, and *Study*, gives a picture of what typical language users can do in a particular language, at a particular level, in one of the skill areas (Listening/Speaking, Writing, Reading.) Table 1 illustrates some typical Can Do statements for different proficiency levels (A1 to C2) in each of the skill and general areas.

Importantly the statements are available in over 16 of the European languages represented in ALTE. For instance, Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish and 'as language-neutral descriptions of levels of language proficiency they constitute a frame of reference to which different language exams at different levels can potentially be related' (Jones and Hirtzel 2001:245.) The statements were developed through a rigorous process involving leading experts on language assessment and applied linguistics and data were collected from 10,000 language learners throughout Europe. The initial calibration of the statements was followed by work on establishing a link between the Can Do scales and the ALTE examinations as well as to the *Common European Framework of Reference for Languages* (CEFR) levels. With the expansion of the membership of ALTE, the Framework is being revised and updated as more data becomes available across the range of languages of the ALTE members. In fact, Peter Hardcastle's article examines the issues involved for ALTE members in comparing tests and the calibration of tests across languages.

Aligning Examinations to the Common European Framework

In addition, ALTE is currently represented on a Council of Europe authoring group which is working on the production of a Pilot Manual for Aligning Language Examinations to the Common European Framework of Reference (2003). The Language Policy Division of the Council of Europe in Strasbourg is developing this in order to assist member states, as well as national and international providers of examinations, in relating their certificates and diplomas to the CEFR. According to the Council of Europe, 'the Manual aims to contribute to competence building in the area of linking assessments to the CEFR; encourage increased transparency on the part of examination providers; encourage the development of both formal and informal national and international networks of institutions and experts' (COE website.) ALTE fully supports this initiative; indeed, some ALTE members are already working on preparing case studies for the manual, while others have given feedback on the pilot manual.

Developing Common Standards and Joint Projects

In addition to establishing a framework of levels on which the examinations can be placed, agreement must be reached on the standards to which these examinations are produced. ALTE members share this concern for establishing common standards for all stages of the language testing process: that is, for test development, task and item writing, test administration, marking and grading, reporting of test results, test analysis and reporting of findings and striving for fairness in the treatment of all candidates who take their exams. In sharing their knowledge and expertise on these issues, they have established a Code of Practice and adopted a Quality Assurance approach based on awareness raising and self-assessment; thereby, recognising the varied linguistic, educational and cultural contexts within which the examinations are being developed and used. It is of enormous value to the thousands of test takers that the tests developed under the ALTE banner are of good quality and that ALTE members are working towards best practice. The results of ALTE's work in this area are freely available on the ALTE website www.alte.org, for instance, in the form of the ALTE Code of Practice and Quality Management Checklists.

Another of ALTE's major aims is to collaborate on joint projects and in the exchange of ideas and know-how. There are numerous instances of such co-operation and space does not permit mention of them all here. However, some examples may serve to indicate the amount of work carried out. For instance, several ALTE representatives formed part of the Working Group set up to represent various professional interests in the development of the *Common European Framework of Reference* document. Moreover, ALTE co-sponsored the development of the Council of Europe

Table 1: Example Can Do statements

	Writing (Work)	**Reading (Study)**	**Listening/Speaking (Social & Tourist)**
C2	CAN make full and accurate notes on all routine meetings.	CAN make full and effective use of dictionaries for productive and receptive purposes.	CAN keep up casual conversations for an extended period of time.
C1	CAN draft a set of straightforward instructions.	CAN assess appropriacy of source material quickly and reliably.	CAN show visitors round and give a detailed description of a place.
B2	CAN write a non-routine letter where this is restricted to matters of fact.	CAN scan texts for relevant information, and grasp main topic of text.	CAN ask for advice and understand the answer, provided this is given in everyday language.
B1	CAN write a simple routine request to a colleague.	CAN assess whether a textbook or article is within the required topic area.	CAN go to a counter service shop and ask for most of what (s)he wants.
A2	CAN leave a simple message giving information.	CAN understand simple visuals on familiar topics.	CAN express an opinion about food.
A1	CAN write a simple routine request to a colleague, such as 'Can I have 20X please?'	CAN read basic notices and instructions.	CAN ask simple questions of a factual nature and understand answers expressed in simple language.

Vantage level specifications, written by Dr John Trim and Professor Jan van Ek (van Ek and Trim 2001) and ALTE members were involved in a project to provide feedback to the authors. Members commissioned expert reviewers, in their organisations or countries, to write reports on Vantage level. A series of regional workshops also took place in various member countries, at which the reports were discussed and additional feedback given to Dr Trim, who attended each workshop. ALTE made the results freely available to the Language Policy Division of the Council of Europe.

ALTE has taken part in the Council of Europe's European Language Portfolio project and has produced, together with EAQUALS (European Association for Quality Language Services) a version of the language portfolio for adult learners. This portfolio (accredited model 06.2000) is now available in various versions including Croatian, English, French, German, Italian and Romanian. A new version in Basque and Spanish is also in preparation and will be published shortly. ALTE and EAQUALS are currently collaborating with the Council of Europe on producing an electronic version of the language portfolio (see Simon Fenn's article).

Co-operation among the ALTE members has produced significant new ALTE products. Organisations from countries in eastern Europe have become interested in developing new testing systems and ALTE members from Estonia, Hungary, Latvia and Lithuania worked together on a Lingua project (submitted in 2003) to extend various previously developed ALTE LINGUA products in the field of test development to Estonian, Hungarian, Latvian and Lithuanian:

- A Glossary of Testing Terms to be used by test developers and teachers (LINGUA reference: 94–09/1801/UK–III)
- Can Do Statements to be used by language learners, users, teachers and testers (94–03/1629/UK–III)
- Guidelines for Test Item Writers to be used by those producing materials for tests (93–09/1326/UK–III)
- Checklists for Test Evaluation to be used by test developers (93–09/1326/UK–III).

A further LINGUA product, the Breakthrough Level (37098–CP–1–96–1–GR–LINGUA–LD) was also extended to Estonian, Latvian and Lithuanian.

A set of descriptors for Breakthrough level was, in fact, first developed by a group of ALTE members (from Finland, Greece, Ireland, Norway and Sweden) who felt that there was a need in their languages for a set of descriptors at a lower level (below the Waystage Level). The group therefore worked on a possible approach to a Breakthrough level, which is now being used as one of the main reference documents for producing a full set of Council of Europe Breakthrough descriptors, a project which ALTE is now concerned with.

ALTE Conferences

Members of ALTE hold 3-day meetings twice a year to discuss projects and plan further work, with each meeting being hosted and chaired by a different member organisation. A development, which first took place in 2000, to include a conference day as part of the meeting, where a wider local audience is invited to come and find out more about the work of ALTE, has proved very successful. At the last such conference day in Krakow in April 2004, presentations were made, by among others, representatives from the European Commission and the Council of Europe, to an audience of over 100 people, on the Common European Framework of Reference in the Polish context; limitations of the Framework in developing comparable examinations and tests; the manual for relating examinations to the CEFR as well as Europe and Languages: The Lingua action of the Socrates programme.

Many of the outcomes of ALTE work are publicly available. For instance, in 2001 an international public conference, which was attended by over 300 delegates, was organised by ALTE in Barcelona to mark the European Year of Languages. The proceedings of this conference were published in 2004 as Studies in Language Testing Vol.18. The second major international ALTE conference will be held in Berlin from 19–21 May 2005. The call for papers is included in this issue on page 24.

Conclusion

ALTE members' shared concern for high standards in language assessment and for fair treatment of the candidates who take their exams has provided a focus for working together and a key objective of the Association. ALTE members are responsible for testing millions of individuals all over the world, though predominantly in Europe, and ALTE is thus in direct contact with enormous numbers of stakeholders: the staff and colleagues working on the exams themselves; professional consultants working on producing and administering the exams, such as item writers and examiners; teachers and other educationalists, including course writers and materials producers; government bodies; test takers and their sponsors (parents, employers); users of test results such as employers and university departments, as well as many others. ALTE' s approach has been and remains an attempt to identify the roles and responsibilities of key stakeholders in the ALTE institutions and to apply a system of quality assurance with flexibility according to the *specific features* of each organisation (i.e. the different ALTE Members and their stakeholder groups). In seeking to establish standards, it is not the aim to make all ALTE Members conform to the same models of assessment for all 24 languages represented, and it is important to recognise the varied linguistic, educational and cultural contexts within which the examinations are being developed and used. An appropriate balance is required between the need to guarantee professional standards to users, and the need to take into account the differing organisational features of the ALTE institutions and the contexts in which their exams are used. Standards are not imposed from "outside" but are established through the mechanism of the system itself and the procedures for monitoring standards are based on awareness raising and self-assessment in the first instance with peer monitoring introduced at a later stage to confirm that the minimum standards are being met. The aim for all ALTE Members is to continue to share expertise and aim at *best practice models* through an on-going process of development.

Cambridge ESOL, as one of ALTE's founding partners, takes

seriously its commitment to improving its own and others' language tests through the work of organisations such as ALTE. Recently Cambridge ESOL staff enjoyed an internal seminar led by Barbara Stevens and Peter Hardcastle which informed staff clearly of the membership and role of ALTE and what challenges lie ahead for this organisation.

ALTE Members

- **Alliance Française, France**
- **Eusko Jaurlaritza** – Department of Education, Universities and Research in the Basque Government.
- **Centre de Langues, Luxembourg**
- **Centre International d'Etudes Pedagogiques, France** – public institute of the French Ministry of Education.
- **Certificaat Nederlands als Vreemde Taal, Leuven** – government-subsidised non-profit organisation for the Dutch as a Foreign Language Certificate.
- **CITOgroep Instituut voor Toetsontwikkeling, Arnhem, the Netherlands** – National Institute for Educational Measurement.
- **Danish Language Testing Consortium** – The Ministry for Refugees, Immigrants and Integration, The University of Copenhagen, Studieskolen in Copenhagen.
- **Generalitat de Catalunya** – The Direcció General de Política Lingüística (DGPL) is the General Directorate of Linguistic Policy in Catalonia.
- **Goethe-Institut, Germany**
- **Instituto Cervantes, Spain** – state organisation responsible for the management of the Diplomas de Español como Lengua Extranjera (DELE).
- **Idegennyelvi Továbbképzõ Központ (ITK) (Centre for Advanced Language Learning, Hungary)** – independent, non-profit institution under the auspices of Eötvös Lóránd University (ELTE).
- **Republic of Latvia, Naturalisation Board** – under the supervision of the Ministry of Justice deals with all issues pertaining to the citizenship of Latvia.
- **Eksami- ja Kvalifikatsioonikeskus (National Examination and Qualification Centre, Estonia)** – the executive organisation of the Ministry of Education and National School Board which organises national exams.
- **Stockholms Universitet, Sweden** – Centre for Research on Bilingualism produces the National Test of Swedish for Adult Immigrants (Sfi-provet) at the request of the National Agency for Education. The Department of Scandinavian Languages in co-operation with the Universities of Gothenburg, Linköping, Lund, Malmö, Umeå and Uppsala is responsible for producing, analysing and developing TISUS (Test In Swedish for University Studies).
- **Studieforbundet Folkeuniversitetet Universitetet i Bergen Norsk Spraktest** – Studieforbundet Folkeuniversitetet and the University of Bergen have been working together under the terms of a collaboration agreement since 1990. In spring 2000, the two institutions formalised their collaboration by creating the Norwegian Language Testing Consortium.
- **Universidad de Salamanca, Spain** – in charge of developing, marking, setting and evaluating the tests leading to the Instituto Cervantes Diplomas de Español como Lengua Extranjera (DELE).
- **Universidad de Lisboa, Portugal** – Departamento de Língua e Cultura Portuguesa (DLCP) is the Department of the University of Lisbon where Portuguese as a Foreign and Second Language is taught.
- **Universita per Stranieri di Perugia, Italy** – University faculty of Italian Language and Culture and special research centres.
- **University of Athens, Greece** – the School of Greek as a Foreign Language of the University of Athens forms part of the Interdepartmental Programme for the Teaching of Greek as a Foreign Language.
- **University of Cambridge ESOL Examinations, UK** – The University of Cambridge ESOL Examinations (Cambridge ESOL) has charitable status as an institution of the University of Cambridge.
- **University of Jyväskylä, Finland** – The National Board of Education and the University of Jyväskylä developed the National Certificate as a joint non-profit project.
- **Uniwersytet Jagiellonski** – The Polish Institute is an interdisciplinary research and teaching centre of the Jagiellonian University in Kraków.
- **Univerza v Ljubljana Slovenia** – The Centre for Slovene as a Second/Foreign Language in the Department of Slavonic Languages and Literatures of the Faculty of Arts, University of Ljubljana.
- **Vilnius University Department of Lithuanian Studies, Lithuania** – Department of Lithuanian Studies, Faculty of Philology of the University of Vilnius.
- **WBT Weiterbildungs-Testsyteme, Germany** – The WBT (Further Education Test Systems) is an institution organised in the form of a limited company under German charity law.
- **Welsh Joint Education Committee, Wales** – WJEC are at present responsible for providing examinations specifically for adults who are learning Welsh as a second language.

Associate Members

- **University of Sofia "St Klimint Ohridski" Department of Language Learning, Bulgaria**
- **Russian Language Testing Consortium: Moscow University and St Petersburg University**

References and further reading

ALTE: http://www.alte.org

ALTE (2004) *Multilingual Glossary of Language Testing Terms* (Latvian) Riga: LR Naturalizacijas parvalde.

ALTE (2004) *Materials for the Guidance of Test Item Writers* (Latvian) Riga: LR Naturalizacijas parvalde.

Council of Europe http://www.coe.int

Council of Europe (2003) *Preliminary Pilot Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment,* Strasbourg: Language Policy Division.

Jones, N and Hirtzel, M (2001) *Appendix D,* in the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe,* Cambridge: Cambridge University Press.

Milanovic, M and Weir, C (Eds) (2004) *European Language Testing in a Global Context, Proceedings of the ALTE Barcelona Conference, Studies in Language Testing Vol.18,* Cambridge: Cambridge University Press.

van Ek, J A and Trim, J L M (2001) *Vantage,* Cambridge: Cambridge University Press.

Test Equivalence and Construct Compatibility across Languages

PETER HARDCASTLE, ALTE VALIDATION OFFICER, RESEARCH AND VALIDATION GROUP

Introduction

Cambridge ESOL is currently involved in validation studies of three suites of European language tests in conjunction with its European partners within ALTE: the University of Salamanca/the Cervantes Institute, The Universitá per Stranieri of Perugia, and the Goethe Institute. All partners either already produce or intend to produce a series of tests aligned to at least five levels of the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001), from A2 Waystage to C2 Mastery. This article reports on an investigation into the constructs, task types and formats of three European language tests of reading at CEFR Threshold Level B1. The tests involved in this study are:

- the Certificati di Conoscenza della Lingua Italiana (CELI); levels 1–5
- the Diplomas de Español como Lengua Extranjera (DELE); levels *inicial, intermedio, superior*
- The Goethe Institute's suite of German tests, including Start Deutsch 1, Start Deutsch 2, Zertifikat Deutsch (and Zertifikat Deutsch für Jugendliche), Zentrale Mittelstufenprüfung (ZMP).

The intention of the project is to link these tests to the CEFR, using established Cambridge ESOL validation procedures, including the application of IRT (Rasch) item calibration and anchoring techniques.

The project sets out to study the equivalence of individual tests over a given time frame examining issues of stability, consistency and reliability, as well as examining test equivalence across languages; for example, did the German B1 level test function at the same CEFR level as the Italian and Spanish B1 level tests in the summer 2003 administrations?

What is the CEFR?

Very broadly speaking, the CEFR is a set of band descriptors/ Can Do statements (around 400), which set out to describe language proficiency in terms general enough to be relevant to all European languages, yet specific enough to provide a basis for generating thought and discussion about test specifications, design and calibration in the context of individual languages. Within ALTE, over the last decade and before, much effort had already been made to develop strategies for comparing tests across languages and the contribution made by this thinking to the development of the CEFR was substantial. The Common European Framework, recognising the importance of this work (see Appendix D of the CEFR – *The ALTE 'Can Do' Statements*), aims to be complete enough to constitute an adequately descriptive model of cross-linguistic language proficiency, yet concise enough to be a practical tool of real use to language teachers and test designers. In a way, however, it is caught between a rock and a hard place, in that its theoretical completeness and hence its descriptive complexity tend to function in indirect proportion to its potential for practical application. Clearly, the framework should encourage, not deter, everyday use, serving as a specialists' and non-specialists' *vade mecum* in the language teaching and language assessment professions. Furthermore (and most importantly for this project), it should be demonstrably and overtly applicable to language testing procedures and language test development. To this end ALTE is developing a Manual and a Code of Practice whose aims are to enable language testers to conceptualise test alignment within the Framework and to consider ways in which justifiable claims, relating tests to CEFR levels, can be made (ALTE 2001, Council of Europe 2003).

As well as its declared potential as a framework 'for the specification of the content of test and examinations' (Council of Europe 2001:18), the CEFR establishes a framework for the description of language proficiency to be applied as benchmark criteria for the definition of proficiency levels across languages; i.e. a B1 threshold level qualification in French is qualitatively equivalent to a B1 in German or Lithuanian etc. It sets out to provide 'a common set of proficiency statements (which) will facilitate comparisons of objectives, levels, materials, tests and achievement in different systems and situations' (ibid. page 16). It is an *implicational scale* employing *band descriptors* which identify the salient linguistic features employed by users of the language at any one of six posited levels of proficiency. The scale is *cumulative* and *hierarchical*, in that descriptors of competences

operationalised at lower levels are viewed as prerequisites for effective performance at the higher levels. In the words of Brian North (2004), one of the CEFR's main authors, 'the descriptor scales can be used for setting objectives, for self-assessment, for teacher assessment and for relating assessment results to common reference points'.

This paper focuses on some of the assessment-related aspects of the Framework.

Issues with the CEFR

A number of issues relating to the CEFR, its robustness, completeness and the applicability of its band descriptors and Can Do statements continue to be raised (Weir 2004). Similarly, Alderson et al. (2004) (also known as the *Dutch Constructs Group*) have pointed out that, although the CEFR claims to be a useful tool for use in the *content specification* of language tests and examinations (ibid. page 19), the Framework is an incomplete instrument for this purpose and needs considerable elaboration before it can be of significant applicability in this area. These limitations impact on the feasibility of this Cambridge ESOL/ALTE project and fall broadly into two categories, relating to *context validity* and *theory-based validity* (see Weir 2004).

Issues relating to context and context validity

- **Task types** are not specified by the framework and the degree to which this impacts on test design and interpretation is significant.
- **Channel** needs to be specified in the descriptors. It is much easier to arrange a meeting face-to-face with a friend than to negotiate a job interview by telephone, yet ability to perform in both of these very different situations may be defined by the same performance descriptor or Can Do statement.
- **The context** in which the task is to be performed is largely unspecified by the framework. It does not adequately take into account theories of cognitive processing, socio-cognitive development and the contextual features of language use (Weir 2004) which lead to deficiencies in its application as a descriptive/comparative tool and to inadequate consideration of elements of authenticity (situational as well as interactional).
- **Can Dos** in one language may involve the application of more sophisticated language skills than in another language, calling into question the feasibility of applying a common set of descriptors to different languages. This is a major issue for the project and is explored further below.
- **How well can they do it?** The level of performance associated with the descriptor is not normally specified or quantified. It is possible to operationalise descriptors at various levels, depending on the complexity of the language used, for example, *Can make and respond to invitations, suggestions, apologies* is a set of functions which could activate complex linguistic exponents, but which could also be performed using simpler language at A2 level, for which it is one of the *Can Do* specifications.

Issues related to theory-based validity

- **Cognitive load** (cognitive processing) is an important component of task design as yet inadequately considered by the CEFR.
- Use of **relativistic** or **value-laden terminology**. Words like *simple, straightforward, short, slow* (speech delivery) etc are frequently used in the descriptors. Such terms are subjective and may be interpreted differently from one user to the next.
- The **terms** employed in the Framework are descriptive and taxonomical rather than developmental and do not reflect an established hierarchical model of cognition or of language development, such as the paradigm developed by Bloom (1956). (Note, however, that the ALTE Can Do statements and the CEFR display an interesting degree of congruence with this early paradigm, at least in terms of the verbs chosen to express the target abilities.)
- It has been suggested that the CEFR does not fit with any current model of language acquisition or theory of language development (Fulcher 2004), though the framework specifically does not set out to apply any such model and incorporates a disclaimer to this effect (ibid. page 18, 2.3.2).

It is not the intention of this article to address these issues at any length, but it is important for any user of the Framework to be aware of these limitations when attempting to employ the Framework for practical ends in language assessment. Without recognition of the limitations, as well as the strengths, of the CEFR it would not be possible to develop a well-founded set of base-references into a fully operational structure for the meaningful comparison of tests.

Cross-language comparability

Cross-language comparability of language tests has to be one of the toughest challenges currently facing the developers of the CEFR and those ALTE members attempting to implement it. Although work has been going on in this domain for several years, much research remains to be done. For the Cambridge ESOL/ALTE project this work is fundamental, as the basic credibility of the CEFR rests on the assumption that comparability of language tests across the levels posited by the Framework can be established by empirical procedures, defined by a set of universal performance descriptors and validated using IRT (Rasch) anchoring techniques. By setting up this project, Cambridge ESOL and its ALTE partners have established an agenda for research into cross-language test alignment based on data collection and analysis. This approach will allow for additional calibration to take place for each set of language tests in their own right (German, Italian and Spanish) using classical and IRT methods. In addition, the question of alignment of the tests to the relevant CEFR levels will be addressed more systematically than in the past. The CEFR performance descriptors, largely in the form of Can Do statements (cf ALTE 2002), have been translated into all of the languages with which this study is currently concerned, but doubt has been expressed as to whether the translated equivalents convey precisely the same meaning.

The concept of *Can Do* in the context of linguistic performance descriptors is one which does not lend itself well to concise and

accurate translation into other European languages. The notion of *ability* is subdivided into more semantic categories in other European languages than it is in English; e.g., two in French (*savoir/pouvoir*), two in German (*können/wissen*), three in Russian (*уметь/знать/мочь*) resulting in sometimes lengthy translations of the Can Do concept, of the order:

- Especificaciones de capacidad lingüística ("puede hacer") (Spanish)
- Kann-Beschreibungen, Kompetenzbeschreibungen (or Fähigkeitsbeschreibungen) (German)
- Indicazioni de capacitá (Italian)
- Capacités de faire (French)

All of the above are less succinct and less exact than the English original.

Similarly, the exponents of the performance descriptors, when expressed in lexico-grammatical terms, may fall into very different categories of difficulty, further complicating the issue of cross-language test calibration based on Can Do statements and/or performance descriptors. For example, the ability to carry out transactions to obtain goods or services (*Can ask people for things and give people things* – level A1) is operationalised at a much higher level of linguistic skill in English when grammatically complex issues of modality and conditionality may arise, than it is in, say, Russian or Arabic where such grammatical complexities do not necessarily occur in the linguistic realisation of this function.

Such incompatibilities are many and varied and present a formidable obstacle to a framework of reference attempting to compartmentalise, describe and calibrate linguistic ability independently of the language in which such abilities are expressed. The comparability tables proposed overleaf hope to go some of the way towards resolving these issues.

Developing Comparability Tables

In response to the CEFR's descriptive framework and in spite of its limitations, this article attempts to construct the modest beginnings of a simplified set of tabulated reference points for the comparison of European language tests, incorporating the elements identified above as omissions from the original Framework. The tests under study are threshold level (B1) measures of reading (written reception) in Spanish, German and Italian, but this concept of horizontal comparability is, of course, applicable to other language tests and at other levels of proficiency. The tabulation is a simplification of test qualities and could be supplemented by many other comparative categories (functional, notional, linguistic, statistical etc.) all of which ultimately form parts of a more exhaustive comparative framework which will need to be produced.

The tables presented overleaf are an attempt to establish the comparability of three language tests in tabular form. They serve only as examples as to how test comparability across languages may be conceptualised. As with the CEFR itself, a disclaimer in this context is more than appropriate: the tables incorporate elements which the Cambridge ESOL/ALTE project has found to be useful in broadly considering test equivalence. The number of quantitative and qualitative fields which could be included in such an instrument are limited only by the imagination of the designer and by the practicalities involved in applying the framework to the analysis of test construct, format and data. The elements included are based on informed judgements and on practical experience and are by no means complete. The three tables contain information under the following headings for one language test each for Italian, Spanish and German:

Key to categories in the comparability tables

1. The *question or section number* in the test.
2. A brief description of the *task type*.
3. *Task format*; e.g. multiple choice (MCQ), multiple matching, gap-filling, live interview.
4. *Number of items* in the task.
5. *Relevant descriptor* refers to the page of the CEFR manual where an appropriate performance descriptor or Can Do statement can be found. Where two descriptors are identified they are both applicable. Where descriptors are identified at two different levels, this should be viewed as a potential difficulty for alignment to the Framework.
6. *Text/discourse type* – an open-ended description of the discourse category into which the text/target language behaviour is best fitted.
7. *PRIM:* production, reception, interaction, mediation (oral or written).
8. *Skill:* reading, writing, speaking or listening.
9. *Sub-skill* is an open-ended sub-division of skill.
10. *General strategy:* components of strategic competence which are mainly paralinguistic or metacognitive behaviours associated with performance on the task (e.g. skimming, scanning, deducing meaning from context).
11. *Construct statement* – users of the table are not expected to enter the full statement into this field, but rather to indicate whether a *construct statement* is available for the information of stakeholders and where this might be found. Developers should be encouraged to produce construct statements for each test and sub-test in order that stakeholders might obtain a full understanding of what is expected of candidates when taking the test, and why.
12. *Register* – various categories, including formal/informal, familiar, legal, journalistic etc, or a combination of these categories.
13. *Task time allocation* – Approximate or exact time allocated to the task.
14. *Linguistic load* – the expected level of difficulty (lexical, grammatical) of the language used (actively or passively) in the task. For tests of spoken language or listening tests, additional aspects of difficulty might be included (speed of delivery, accent etc). This category has to be judged by largely intuitive

criteria. In itself it is, of course, subjective, but when viewed in conjunction with other categories it may help to identify discrepancies or anomalies. (Clearly, a more precise scale will have to be developed in association with this field. Such a scale could be six-point Likert-type scale, compatible with the six levels of the CEFR and could be developed along the lines of the work of Van Ek and Trim (1998, 1999, 2001), referring specifically to their level-specific grammatical summaries and to the lexical indices.)

15. *Cognitive load* – non-linguistic elements of information processing which may affect the way the task is performed. An example is a writing task which uses an *information transfer* input prompt requiring non-linguistic interpretational skills and an element of world-knowledge, on which adequate completion of the task may depend. Such a question would have a high cognitive load. Again, this is a category whose values must be established by intuitive criteria, based on experience. A more refined scale than the one used here (low-medium-high) is under development.
16. *Lexical load* – a quantitative category which considers the approximate number of lexical items to be processed per answer (the lexical input: response ratio). This identifies the number of words which must be read/heard/understood/written for each point earned. In the tests investigated, variation in this area is surprisingly large. Care must be taken when word counts for lexically agglutinative languages (e.g. German) are interpreted.
17. *Situational authenticity* – a judgement of the degree to which the task is representative of a real-life activity. Again, a more refined scale than the one used here (low-medium-high) is under development.
18. *Interactional authenticity* – a judgement of the extent to which the task is concomitant with a real-life language processing task. A task could have high situational authenticity but low interactional authenticity, or indeed vice-versa (see Bachman 1990).
19. *Rasch difficulty range (target)* using Cambridge ESOL scaled logits. Raw logit scores could equally well be used, provided that they were related to a global difficulty scale established empirically by Rasch calibration techniques.
20. *Rasch difficulty range (established)* – the difficulty of task and test after calibration.

A comparison of the three tables (even though they refer only to one skill in only three of the dozens of European languages tests currently available) begins to address some of the issues outlined earlier in this article. The categories in the table are not exhaustive and it would not be difficult to include an additional twenty categories and remain incomplete. Clearly, however, for such an instrument to be useful it must trade off the comprehensive against the practical and reach a useful compromise. From a review of the information in the tables a degree of compatibility among the three language tests can be observed, though blank spaces in the grid indicate that information remains to be collected and inserted.

One of the main absences relating to all three tests is that of the Rasch calibration results (column 20) on which this project is currently working. These results will shortly be entered and reviewed along with the other test properties relevant to calibration and comparability.

Conclusion

Although at an early stage of development this basis for the comparison of tests across languages shows promise, especially when used in conjunction with the CEFR and the Rasch anchoring techniques used by Cambridge ESOL. More work needs to be done, more categories (both quantitative and qualitative) need to be established, but as a rough guide to test equivalence the tables used in this study are already proving useful in establishing tentative equivalence among European language tests.

References and further reading

Alderson, J C, Figueras, N, Kuijper, H, Nold, G, Takala, S, Tardieu, C (2004) *Specification for Item Development and Classification within the CEF; Reading and Listening (English, French and German): The Dutch CEF Construct Project*, paper presented at a workshop to discuss research into and with the Common European Framework, University of Amsterdam.

ALTE (2001) *Principles of Good Practice for ALTE exams*, Revised Draft, Oct 2001, ALTE.

ALTE (2002) *The ALTE Can Do Project*, Association of Language Testers in Europe.

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

Bloom, B S and Krathwohl, D R (1956) *Taxonomy of Educational Objectives: The Classification of Educational Goals, by a committee of college and university examiners. Handbook I: Cognitive Domain*, New York: Longmans, See: http://faculty.washington.edu/krumme/guides/bloom.html

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

Council of Europe (2003) *Preliminary Pilot Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Strasbourg: Language Policy Division.

Fulcher, G (2004) Are Europe's tests being built on an 'unsafe' framework? *Education Guardian*. Retrieved from: http://education.guardian.co.uk/tefl/story/0,5500,1170569,00.html

North, B (2004) Europe's framework promotes language discussion, not directives. *Education Guardian*. Retrieved from: http://education.guardian.co.uk/tefl/story/0,5500,1191130,00.html

Taylor, L (2004) Issues of Test Comparability, *Research Notes* 15, Cambridge: Cambridge ESOL.

van Ek, J A and Trim J L M (1998) *Threshold 1990*, Cambridge: Cambridge University Press.

—(1999) *Waystage 1990*, Cambridge: Cambridge University Press.

—(2001) *Vantage*, Cambridge: Cambridge University Press.

Weir, C J (2004) *Limitations of the Common European Framework of Reference (CEFR) in Developing Comparable Examinations and Tests*, paper delivered at ALTE conference, Kraków, April 2004.

—(forthcoming) *Language Testing and Validity Evidence*, Basingstoke: Palgrave Macmillan.

TABLE 1: TEST COMPARABILITY TABLE FOR READING COMPREHENSION TASK TYPES IN DELE

Test: *DELE Level 1 (Nivel Inicial)* ***CEFR Level:*** *Threshold (B1)* ***Language:*** *Spanish* ***Target Skill/s:*** *Reading* ***Total Time:*** *2 hrs*

1	2	3	4	5	6	7	8	9	10
Question	**Task Type**	**Format**	**No. items**	**Relevant descriptor (CEFR)**	**Text/discourse type**	**PRIM (o/w)**	**Skill**	**Sub-skill**	**General strategy**
1	Reading comp. passage	MCQ (4 options)	5	p70 B1(ii)	Tourist information (descriptive/ informational)	r (w)	r	Extracting main point from text	Intensive reading for info.
2	Adverts (c. 30 words)	Matching (5 with 7)	5	p70 A2(i) & B1(ii)	Accommodation information (informational)	r (w)	r	Assessing validity of text/related statement	Skimming/ scanning
3	Adverts (c. 60 words)	MCQ (4 options)	5	p70 A2(i) & B1(ii)	Publicity (suasive)	r (w)	r	?	Skimming
4	Adverts (c. 60 words)	4 texts – matching exercise	10	p70 A2(i) & B1(ii)	Publicity/personal announcements, recipes etc.	r (w)	r	?	Skimming
5	Sentences (c. 30 words)	Sentence completion by matching	10	?	Tourist information (descriptive/ informational)	r (w)	r	Grammatical competence (pronouns)	Identifying grammatical features

TABLE 2: TEST COMPARABILITY TABLE FOR READING COMPREHENSION TASK TYPES IN CELI

Test: *CELI (Certificato de Conoscenza della Lingua Italiana) Level 2* ***CEFR Level:*** *Threshold (B1)* ***Language:*** *Italian* ***Target Skill/s:*** *Reading* ***Total Time:*** *2 hrs*

1	2	3	4	5	6	7	8	9	10
Question	**Task Type**	**Format**	**No. items**	**Relevant descriptor (CEFR)**	**Text/discourse type**	**PRIM (o/w)**	**Skill**	**Sub-skill**	**General strategy**
A1	7 short texts	MCQ (3 options) 1 item per text	7	p70 B1	Informational	r (w)	reading	Extracting main point from text	Identifying main points (skimmimg)
A2	Reading text (c. 300 words)	10 questions (true/false)	10	p70 B1	Comparative	r (w)	reading	Assessing validity of text/related statement	Evaluating validity of a proposition
A3	Sentence completion	MCQ (4 options)	5	?	Short individual sentences	r (w)	reading	?	Intensive reading
A4	Selective deletion cloze	MCQ (3 options)	10	p69 B1(i)	150 word passage	r (w)	reading	?	Intensive reading
A5	Sentence completion	gap filling (using pronouns)	5	?	Short individual sentences	r/p	reading/ grammar-based skills	Grammatical competence (pronouns)	?

TABLE 3: TEST COMPARABILITY TABLE FOR READING COMPREHENSION TASK TYPES IN ZERT

Test: *ZERT (Zertificat Deutsch)* ***CEFR Level:*** *Threshold (B1)* ***Language:*** *German* ***Target Skill/s:*** *Reading* ***Total Time:*** *90 minutes*

1	2	3	4	5	6	7	8	9	10
Question	**Task Type**	**Format**	**No. items**	**Relevant descriptor (CEFR)**	**Text/discourse type**	**PRIM (o/w)**	**Skill**	**Sub-skill**	**General strategy**
1	5 short (130 word) Reading passages	Matching	5 answers (choice from 10)	p70 B1(ii) & B1(iv)	Tourist information (Informational)	r (w)	reading	Extracting main point from text	Identifying main point
2	Extensive Reading passage (400 words)	MCQ	5	p70 B1(i) & (iv)	Discussion	r (w)	reading	Assessing validity of text/related statement	
3	Adverts (12)	Matching	10 (choice from 12)	p70 B1(ii), A2(i), B1(iv)	Suasion, information	r (w)	reading	?	Scanning
4	Selective deletion cloze	MCQ (3 options)	10	?	Informational	r (w)	reading (with lang. use)	?	
5	Selective deletion cloze	Matching	10 (choice from 16)	?	Information, suasion (publicity letter)	r (w)	reading (with lang. use)	?	

General Construct Statement: *To be developed*

11	12	13	14	15	16	17	18	19	20
Construct statement	**Register**	**Task time allocation**	**Linguistic load**	**Cognitive load**	**Lexical load**	**Situational authenticity**	**Interactional authenticity**	**Rasch difficulty range (target)**	**Rasch difficulty range (actual)**
To be developed	Formal	25 mins	Fairly high	Low	100:1	High	Low	48–68	To be established
	Informal/ publicity	25 mins	Low	Low	100:1	High	Low	48–68	
	Informal/ publicity	20 mins	Low	Low	90:1	High	Low	48–68	
	Informal/ publicity	20 mins	Low	Low	70:1	High	Low	48–68	
	Formal	15 mins	Fairly high	Low	25:1	Low	Low	48–68	

General Construct Statement: *To be developed*

11	12	13	14	15	16	17	18	19	20
Construct statement	**Register**	**Task time allocation**	**Linguistic load**	**Cognitive load**	**Lexical load**	**Situational authenticity**	**Interactional authenticity**	**Rasch difficulty range (target)**	**Rasch difficulty range (actual)**
To be developed	Formal/ informal journalistic	20 mins	Fairly high	Medium	150:1	Medium	Medium to Low	48–68	To be established
	Semi-formal	15 mins	Fairly high	Medium	30:1	Medium	Low	48–68	
	Semi-formal	5 mins	Low	Low	10:1	Low	Low	48–68	
	Semi-formal journalistic	10 mins	Medium	Low	15:1	Low	Low	48–68	
	Formal	5 mins	Low	Low	10:1	Low	Low	48–68	

General Construct Statement: *To be developed*

11	12	13	14	15	16	17	18	19	20
Construct statement	**Register**	**Task time allocation**	**Linguistic load**	**Cognitive load**	**Lexical load**	**Situational authenticity**	**Interactional authenticity**	**Rasch difficulty range (target)**	**Rasch difficulty range (actual)**
To be developed	Journalistic	20 mins	Fairly high	Medium	130:1	Medium	Low	48–68	To be established
	Journalistic	15 mins	Fairly high	Medium	100:1	High	Low	48–68	
	Informal/ publicity	5 mins	Low	Low	80:1	High	Low	48–68	
	Informal/ personal	10 mins	Medium	Medium	20:1	Medium	Low	48–68	
	Informal/ personal	10 mins	Medium	Medium	20:1	Medium	Low	48–68	

Development of an Electronic European Language Portfolio

SIMON FENN, ESOL PROJECTS OFFICE, OPERATIONS GROUP

The European Language Portfolio (ELP) was developed and piloted from 1998 until 2000 by the Language Policy Division of the Council of Europe, Strasbourg. It was launched on a pan-European level during the European Year of Languages, 2001 as a tool to support the development of plurilingualism and pluriculturalism.

The ELP is based on the six level system of the Council of Europe's Common European Framework of Reference for Languages (A1, A2, B1, B2, C1, C2) and includes the overall grid of descriptors by level and skill from the CEFR. It consists of the following elements:

- **Language passport**
  'An updateable overview of your experience in and ability with different languages. It records formal qualifications and diplomas, and self-assessments.'
- **Language biography**
  'A record of your personal language learning history which helps you to evaluate your learning objectives, and reflect on your language learning and inter-cultural experiences.'
- **Dossier**
  'A collection of pieces of work and certificates chosen by you to document and illustrate your language skills, experiences and achievements.'

Language portfolios are used by language learners to record and reflect on their language learning achievements and aspirations and to demonstrate to others their language skills and their knowledge of other cultures. The ALTE/EAQUALS validated European Language Portfolio for adults was introduced in 2002 in four languages French, German, English and Italian and has been further developed in Croatian, Romanian, Basque and Spanish.

Now, moving beyond this paper-based ELP, ALTE and EAQUALS, in collaboration with Council of Europe, have begun development of an electronic version, known as the ALTE/EAQUALS electronic European Language Portfolio or eELP. The eELP will be a downloadable application which will be installed locally on computers. It will be built using Java, a programming language which allows the application to run on a variety of platforms and operating systems. Anyone wishing to use the new eELP will be able to download it from a website. The eELP will be available free of charge but users will be required to register their details before they are able to download and install the software. Once installed, the application will support multiple users, an important consideration if the eELP is to be used on shared computers in language schools or other organisations. Users will be able to login to the application and access a number of functional and content-based areas.

The Language Passport contains a number of areas, some of which are listed below:

- The *Profile of Language Skills* allows language learners to determine and record their language learning level by completing self assessment grids, the levels of which map to the levels in the CEFR. Being able to do this on screen, rather than on paper, is more effective for language learners as they are able to control how many grids are available according to the number of languages the learner wishes to record. Cutting down on the 'white space' of the paper ELP is seen as a key objective in the development of the eELP.
- The *Certificates and Diplomas* section allows the language learner to record the details of any language examinations awarded.
- The ALTE framework of examinations and grids to fill in showing courses taken in EAQUALS schools.
- *Content areas*, explaining how the language learner should complete the Language Passport.

The other main area of the eELP is the *Language Biography*. While the Language Passport concentrates on recording language learners' achievements, the Language Biography concentrates on why language learners are learning a particular language and their aims and aspirations, in practical terms, for learning the language. The Biography helps language learners to structure their language learning experiences in order that they can assess how best to progress their language learning. The Biography will consist of the following areas, essentially a replication of the paper ELP but enhanced to make best use of the electronic medium:

- *My Language Learning Aims* allows language learners to add specific aims which they wish to achieve in particular languages. Once again the electronic format is more user friendly so instead of presenting new language learners with a large and rather intimidating blank grid, the language learner can add in new aims, as and when required.
- Similar sections will exist to allow language learners to record their *Language Learning History*, *Significant Linguistic and Intercultural Experiences* and *Current Language Learning Priorities* allowing language learners to build up a wide-ranging and detailed biography of any language of which they have any experience.
- The *Dossier* will provide language learners with functionality to add pieces of work or details of achievements, aimed at demonstrating and acting as evidence of language learning achievements. Where the piece of work exists in an electronic format, such as an audio file or Word document, it can be stored in the eELP. Where the work or achievement cannot be stored electronically, for example where the language learner wants to reference a certificate or diploma, details can be stored but not the item itself (although the item could be scanned).

In addition, the eELP will have a *Roadmap*, which is essentially an online help facility, guiding language learners through the different areas of the eELP.

The project to build the eELP is being supported by the Cambridge ESOL Projects Office. The eELP must conform to the Principles and Guidelines for European Language Portfolio models and be validated by the Validation Committee of the Council of Europe, the next meeting of which is in November 2004, meaning tight deadlines for the development of this product. Detailed design of the new application started in June 2004, following a high-level requirements gathering phase. The detailed design phase will be followed by an intensive build phase where developers will produce the new eELP, which will then be tested by ALTE and EAQUALS. The new product will be available to users to download in January 2005, following successful validation and testing. The eELP will initially support English and French as the two languages of administration, although future versions will support additional languages. Other suggested future enhancements include making the eELP interact with web based functionality which would mean that language learners would be able to share their achievements more easily.

Cambridge ESOL supports the eELP project as it recognises the wider benefits to individuals and societies that access to, and achievement in, language learning opportunities bring.

For further information please see the ALTE website: www.alte.org or contact the Secretariat: stevens.b@ucles.org.uk

Automated Writing Assessment: a review of four conceptual models

STUART D SHAW, RESEARCH AND VALIDATION GROUP

Introduction

With the advent of a burgeoning testing candidature, the increased demand placed upon large-scale assessment programmes in evaluating responses to direct tests of writing is widely recognised. Direct tests of writing – essay writing designed to measure writing ability, along with open-ended questions requiring short answers – have long been highly valued components of effective assessment programmes.

UCLES is investigating the feasibility of automatically grading essays using a computer. In this context an essay is simply a textual response to a question typed into a computer including any number of words from a few sentences upward. The aim of these studies is to investigate a number of possible approaches to the automatic assessment of natural language essays. Human assigned grades represent an overall judgement of the quality of language form, structure and content in the essay. The research question raised by such studies is whether automatic techniques can simulate human judgement.

Clearly, the practical implications of either automated or partially-automated rating are obvious – not simply in terms of increasing the speed of turning around results or the money saved but also in terms of the automatic generation of valuable diagnostic feedback which could be of use to learners, teachers and testers alike.

UCLES' primary interest in automated writing evaluation of 'free' or extended text was engendered by the development and implementation of computerised systems capable of assessing essays automatically. This article provides a review of the theoretical models of four of the most prominent implemented systems, and attempts to appraise their respective strengths and weaknesses.

Automated assessment

Traditionally, open-ended items have been widely perceived to be unsuitable for machine marking because of the difficulty of handling the multiform ways in which credit worthy responses may be expressed. The extent of successful automatic marking of free text answers would seem to presuppose, at least in part, a sophisticated level of performance in automated natural language understanding. More recently, however, advances in Natural Language Processing (NLP) techniques have revealed potential for automatic assessment of free text responses keyed into a computer without the need of having to create system software that fully comprehend the responses. In essence, NLP is the application of computational methods to analyse characteristics of electronic files of text (or speech). NLP applications utilise tools such as syntactic parsers which analyse the discourse structure/organisation of a response (Marcu 2000) and lexical similarity measures which analyse word-use of a text (Salton 1989).

Perceptions of automated assessment

Despite understandable scepticism regarding the potential value of automated essay assessment (Wresch 1993), some reactions have been decidedly favourable. Apart from being cost effective, computerised scoring is unfailingly consistent, highly objective and almost wholly impartial (Schwartz 1998). There are those, however, who regard the concept of computer-assisted evaluation as being incompatible with current notions of communicative writing proficiency, which stress – amongst other things – the writer's ability to communicate, or engage, a specific readership.

It is the view of some critics that, unlike human raters, computers are incapable of differentiating between inspirational and creatively exceptional essays and their technically correct but ordinary

counterparts (DeLoughry 1995; Mitchell 1998). This recognised shortcoming originates, it would seem, from the presumption that automated scoring emphasises linguistic rules and grammatical conventions at the expense of less tangible or demonstrable qualities, such as textual clarity and overall coherence. In accordance with this view, computers may be able to investigate writing for the presence or the absence of certain words, phrases or structures, but they cannot be expected to appreciate a writer's communicative purpose in the same way that human raters can. Even the developers of automated systems readily acknowledge that – although such systems can be useful tools in assessment – they are unable to adequately replace effective writing assessors.

Four conceptual models for automated essay assessment

1. *Project Essay Grader* (PEG) can trace its beginnings to work undertaken during the 1960s by Page (1966) and is widely regarded as the first and most enduring implementation of automated writing evaluation. It depends, principally, on the linguistic features of an essay;
2. *E-rater* (or Essay-rater) is a system developed by the Education Testing Service (ETS) (Burstein et al. 1998a; Burstein and Kukich 1998a; Burstein, Leacock and Swartz 2001). The system – which uses shallow parsing techniques to identify syntactic and discourse features – employs a hybrid approach of combining both statistical tools and linguistic features derived by using NLP techniques to model the decision of a human assessor. An additional but related strand of research activity currently being undertaken by ETS Technologies is an investigation into the feasibility of automating the rating of short answer content-based questions. The *C-rater* prototype, again utilising NLP technology, is being evaluated for its effectiveness at producing 'credit/no credit' ratings;
3. *Latent Semantic Analysis* (LSA) often referred to as the 'bag of words' approach. LSA was developed by Landauer (University of Colorado) and is based on word document co-occurrence statistics represented as a matrix, which is subsequently decomposed and then subjected to a dimensionality reduction technique. Latent Semantic Analysis is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. LSA is a machine-learning technology for simulating the meaning of words, phrases and stretches of textual discourse and effectively ignores essay linguistic and structure features i.e. the LSA technique evaluates content via the choice of words and does not take into account any syntactic information;
4. *Text Categorisation Techniques* (TCT) – developed by Larkey (University of Massachusetts, 1998) – utilise a combination of modified key words and linguistic features. Text categorisation is the process of grouping text documents into one or more predefined categories based on their content. Several machine learning methods and statistical classifications have been applied to text categorisation including Bayesian classifiers, nearest neighbour classifiers, decision trees, neural networks and support vector machines.

Project Essay Grader model

Acknowledging the enormous rating load placed upon teachers, trainers and large-scale assessment programmes in evaluating student essays, Page developed an automated essay rating system. Beginning with an established corpus of pre-rated student compositions, by experimenting with a combination of automatically extractable textual features, Page applied multiple linear regression techniques to ascertain an optimal permutation of weighted features that most accurately predicted the teachers' ratings. The system was then able to rate – by adopting the same set of identified weighted features – other student essays. In their developmental work, Page and his research associates adopted two explanatory terms:

> *'Trins were the intrinsic variables of interest – fluency, diction, grammar, punctuation, and many others. We had no direct measures of these, so began with substitutes: Proxes were approximations, or possible correlates, of these trins. All the computer variables (the actual counts in the essays) were proxes. For example, the trin of fluency was correlated with the prox of the number of words.'* (Page 1994:130)

Multiple regression techniques employed by the PEG model are used in the computation, derived from the proxes, of an algorithm to predict a rating for each essay.

Evaluation of PEG

Page's system demonstrated an inter-rater multiple correlational index with 'human' scores of 0.78 – nearly as high as the claimed 0.85 correlation between two or more human raters. Despite its undoubted early success, initial versions of PEG received merely partial recognition amongst members of the educational community primarily due to the fact that it only addressed *indirect* measures of writing skill (because of the computational difficulty of implementing more direct measures). Indirect measures of writing ability provided students with the potential for cheating as it was possible to affect the rating by artificially inflating scores through the use of various strategies (for example writing lengthier compositions) and other unfair methods. Indirect measures could not capture certain rhetorical, organisational and stylistic writing features which meant that PEG was incapable of generating diagnostic information, a type of feedback valued by both teachers and students.

In 1995, PEG underwent a series of radical changes designed to embrace more direct measures of writing ability. Despite some success, the principal research challenge must continue to be one of identifying and automatically extracting more *direct* measures of writing quality.

E-Rater model

E-rater employs a combination of statistical and NLP techniques to extract linguistic features of the essays to be rated. It has been used to rate GMAT (*Graduate Management Admissions Test*) and TWE (*Test of Written English*) essays for prospective tertiary-level students.

Essay-rating specifications for the *Analytical Writing Assessment*

portion of the GMAT specify a set of general writing assessment characteristics: topic content, syntactic variety and organisation of ideas. A pilot version of the computerised *Analytical Writing Assessment* enabled the generation of data necessary for a preliminary automated essay rating investigation to be undertaken. Researchers at ETS identified and defined over 100 automatically extractable essay features and computerised algorithms were employed to extract values for every feature from each essay. Stepwise linear regression methods were applied to subsets of features in order to ascertain optimised rating models – models which were predictive of ratings awarded by 'expert' assessors.

E-rater comprises modules that automatically extract essay lexical content, discourse structure and syntactic (grammatical structure) information. Multiple linear regression techniques are subsequently employed to suggest a rating for the essay, based upon the very features extracted by the model.

Subsequent refinements to the system have been implemented by ETS (Burstein et al. 1998a, 1998b; Burstein, Leacock and Swartz 2001). Shallow parsing techniques, employed ostensibly to identify syntactic and discourse features, are fundamental to the model. Content is checked by vectors of weighted content words. An essay that remains focused i.e. stays 'on-track', is coherent as evidenced by use of discourse structures and demonstrates good lexical resource and varied syntactic structure is awarded a higher score.

Additionally, ETS have researched the feasibility of automating the rating of short answer content-based questions and have developed an automated rating prototype called *C-rater* which uses NLP technology.

Evaluation of E-rater

E-rater has been evaluated by Burstein et al. (1998a) who have observed levels of agreement with human raters of between 87% and 94% – on a par with that found between human raters. Empirically founded, E-rater is 'trained' through exposure to examples of 'gold standard' essays (previously rated by human readers) – the samples being selected according to essay prompt so as to ensure they exemplify a comprehensive range of available scores. Although empirically based, E-rater is not 'blindly' empirical in the sense that it could readily adopt indiscriminate selection criteria for extracting various features. However, available extractable features do not always correspond equally well nor do they contribute uniformly to the features that human raters are trained to address when rating. E-rater features are, as a consequence, required not only to be predictive of human ratings, but also to demonstrate a degree of logical correspondence to the various characteristics that human raters are trained to consider. These characteristics are made explicit in the available rater training documentation and/or the accompanying rubrics.

Results from an exploratory, small-scale C-Rater study with a university virtual learning programme have been extremely promising with C-rater achieving over 80% agreement with the score awarded by an expert. This area of research has the potential to evolve into an automated scoring application that would be appropriate for evaluating user-constructed responses in 'on-line' assessment applications.

Latent Semantic Analysis model

LSA seeks to look beneath the surface lexical content of an essay to quantify its deeper semantic content. The fundamental principle underpinning LSA is that the aggregate of all the contexts in which words or lexical chunks occur provides a huge complex of simultaneous equations that determine the similarity of meaning of various words and portions of text to each other. Words and their various contexts are represented by a large two-dimensional matrix. A word is represented as a row whilst a column represents larger portions of context (sentences, paragraphs) in which the words are to be found. By employing a matrix algebra technique known as *Singular Value Decomposition* (SVD) – deconstructing the matrix into three component matrices and then subjecting the matrices to a dimensionality reduction technique – new relationships between words are revealed and already existing relationships are changed in order for their true importance to be more accurately represented.

In order to rate an essay, LSA first constructs an essay matrix. The matrix is subsequently transformed by the SVD algebraic technique to replicate approximately the matrix using the reduced dimensional matrices constructed for the essay topic domain semantic space. Each word and sentence is represented in LSA as a discrete point in a multi-dimensional semantic space (the semantic space typically consists of human rated essays). Similarity of meaning estimates between any two words are derived from their relative positions in this semantic space. Vectors are then calculated from data contained within the student's essay and these essay vectors and all essays contained within the semantic space are compared against each other. Finally, the award for the rated essay with the lowest cosine value in relation to the essay to be rated is assigned.

The Intelligent Essay Assessor (IEA) – developed by Knowledge Analysis Technologies (KAT), Colorado (Foltz, Laham and Landauer, 2003) – uses LSA for rating essay responses and is used to make comparisons between students' responses and 'exemplar' responses through computation of the 'distance' between their corresponding vector projections (Graesser et al. 2000). IEA has been subject to a variety of testing scenarios: essay comparison with pre-scored ratings, comparison with *gold standard* essays (Wolfe et al. 1998), to sections of the original text, or to sub-components of texts or essays (Foltz 1996; Foltz, Britt and Perfetti 1996).

Evaluation of LSA

Foltz (1996) observes that LSA rating performance is almost as reliable as human assessors. Percentages for adjacent agreement with human raters fall in the range 85%–91% on GMAT essays.

The LSA technique evaluates content via the choice of words and is unable to account for syntactic information. In this sense it is a 'bag of words' approach and as such can be tricked as the following quote graphically demonstrates:

> *'It has no way of knowing the difference between The Germans bombed the British cities and The British bombed the German cities … The machine would grade both these sentences the same way.'* (Charles Perfetti 1998)

Nor can LSA adequately cope with phenomena considered difficult for NLP systems, like negation, attachment, binding, predication, modification, scope ambiguities and so on. Attempts have been made to enhance the performance of LSA by incorporating certain syntactic and semantic information: adding a part-of-speech to the given word (Wiemar-Hastings and Zipitria 2001) or adding a part-of-speech to the previous word (Kanejiya, Kumar and Prasad 2003). Such enhancements have not, however, produced a significant improvement over the basic method.

Text Categorisation Technique model

From one perspective, LSA is simply a way of categorising a text. Other researchers (Rosé et al. 2003; Larkey 1998) have employed text-categorisation (or classification) for the rating of qualitative physics questions and law responses (where the text is expected to comprise a legal argument). While the accuracy of such systems may not be able to exceed that of hand-crafted systems, they nevertheless have the advantage of being automatically customisable to new domains requiring no other expert knowledge than that of a human assessor.

Basing his model on Text Categorisation Techniques, Larkey's approach to automatic assessment is to:

> *'… train binary classifiers to distinguish 'good' from 'bad' essays, and use the scores output by the classifiers to rank essays and assign grades to them'.* (1998:90)

TCT depends, primarily, on the analysis of the occurrence of particular words in documents. The approach makes use of Bayesian independent classifiers (Maron 1961) for assigning probability indices to documents, the computation of membership likelihood estimates and the adoption of *k nearest neighbour techniques* (Mitchell 1997).

K nearest neighbour techniques start with a set or range of exemplars representing the main possible outcomes (k is deduced through training the automated system on a set of human rated essays). When this particular part of the training phase has been completed, new examples requiring analysis are matched with the established exemplars (or combinations of them) and the closest match determines the appropriate response. In regard to essay rating, the 'analyses' are ratings and the rating of the nearest matching exemplar is assigned to the input text to be rated.

New input text (examples) are decomposed into a set of *features* (words, grammatical relations, synonym sets and so on). By beginning with word tokens, say, it is possible to approximate a crude rating method for identifying keywords in responses. A weight – known as the 'tf-idf' measure – is assigned to each content word which appears in the training set. The 'tf-idf' measure is *term frequency* (the number of times the term or feature appears in the example) multiplied by *inverse document frequency*, i.e. 1/(the number of times term or feature appears in all examples). Clearly, terms which do not distinguish well among examples carry less weight. It is possible, therefore, to assign higher weights to words that are associated with (in)correct responses, and less weight to words that occur in almost all responses. Exemplars and any new input can be represented as a vector of weighted feature values (i.e. linguistic properties believed relevant), ordered and categorised in some canonical way. It is then possible to compute a cosine or similar distance measure between the training exemplars and the input text to be rated.

Evaluation of TCT

K nearest neighbour results derived from preliminary work undertaken by UCLES using information extraction and retrieval techniques to mark General Certificate of Secondary Education (GCSE) biology answers (described below) reveal that on average the number of times an automatic rating matches the examiner's rating is 67 in 100 i.e. 67%.

Larkey (1998) – in his linear regression trials – automatically rated general opinion compositions (in which style constituted the main assessment criterion). The system performed well: the proportion of compositions automatically rated exactly the same as those rated by humans was of the order of 0.6. (0.97 for ratings within +/– one grade of human ratings).

Comparison of automatic essay assessment models

The table below depicts – in summary form – several characteristics of the four automated writing assessment models described and shows reported model performances (in comparison to human assessors), shared variation between model-human and human-human ratings, and the particular approaches employed by each of the four models. As is the case for each of the models discussed here, essays are evaluated against a set of 'gold standard' scripts i.e. a benchmark set of human rated essays.

Table 1: Comparative performance of automated assessment models

Model	Performance * % ** correlation	Shared variation (as a %)	Approach
E-rater	87–94*	87–94	hybrid – linguistic features (NLP) + document structure features
LSA	85–91*	85–91	"bag of words" – ignores document linguistic + structure features
TCT	0.60–0.80**	36–64	combination of modified key words + linguistic features
PEG	0.39–0.78**	15–61	linguistic features

The most sensible way of interpreting the correlation coefficients shown in the second column is to convert them into overlap between the two measures. To compute the overlap, the square of the correlation coefficient is taken. This enables us to see how much of the variance in one measure can be accounted for by the other. To the degree that the two measures correlate, they share variance. Thus PEG performance accounts for between 15% and 61% of the variations between PEG and human ratings, whilst TCT accounts for between 36% and 64%. It would appear that E-rater tends to perform marginally better than LSA and very much better than either TCT or PEG.

Automatic marking of short textual answers using information extraction

A highly regarded feature of many of UCLES examinations (including Cambridge ESOL tests) is the widespread use of questions that require – from candidates – one or two sentences as a response. Any system that could either partially or completely automate the valid marking of short, free text answers would, therefore, be of great value. Until comparatively recently this has been considered either impossible or impractical. However, recent innovations in computational linguistics together with an increasing emergence of computers in the classroom, have triggered a number of assessment organisations to explore the possibility of automatic marking and its application to high or low-stakes tests. UCLES is one such organisation. The *Interactive Technologies and Assessment Learning Unit* (ITAL) is funding a three-year study at Oxford University which began in summer 2002. The project employs information extraction and retrieval techniques to mark GCSE biology answers using an automatic short answer assessor called *Automark 3*.

Automark 3 has been developed by *Intelligent Assessment Technologies* (Mitchell et al. 2002) and employs information extraction techniques in the sense that the content of a correct response is specified in the form of a number of mark scheme templates. The stretch of text to be rated is fed into a parser (in this case the Link Grammar parser (Sleator and Temperley 1991, 1993)) and the resulting parsed text is then compared to the already-defined templates or mark scheme. Mitchell et al. (2002) claim about 95% agreement with human markers in blind testing. Callear, Jerrams-Smith and Soh (2001) at the University of Portsmouth also use pattern-matching techniques to mark short answers in programming languages, psychology and biology – related fields.

Preliminary findings demonstrate that information extraction techniques can be successfully employed for the task of marking GCSE biology scripts. It has also been shown that a relatively naïve text classification method can rate better than a simple baseline grading technique. There are still many refinements to the approach that can be usefully made: the final aim is to attempt to approach the accuracy of the information extraction method but using completely automatic machine learning techniques.

Conclusion

The potential for using automatic essay scoring in operational testing environments reduces both the time and the significant costs associated with having multiple human raters manually assess essay responses. The agreement between two human raters, and between the conceptual models presented in this article and a human rater is very favourable. Automated essay scoring would, therefore, appear to be a very feasible solution toward the introduction of more writing assessments in certain high-stakes standardised tests, and in a lower stakes context, for example for the purposes of classroom instruction. Moreover, the increased availability of these technologies may well provide incentives for making a greater quantity and range of assessment and instructional materials available 'online'.

In the final analysis, however, human raters will always be required to work alongside automated systems, for example in preparing exemplar responses and to rate samples of student essays before the assessment systems described here are able to complete the task.

References and further reading

Burstein, J and Kukich, K, (1998) Computer analysis of essays, in Proceedings of the NCME Symposium on Automated Scoring, April 1998, Montreal, Canada.

Burstein, J, Kukich, K, Wolff, S, Chi, Lu and Chodorow, M (1998a) Enriching automated scoring using discourse marking, paper presented at the Workshop on Discourse Relations and Discourse Marking conducted at the Annual Meeting of the Association of Computational Linguistics, Montreal, Canada.

Burstein, J, Kukich, K, Wolff, S, Chi, Lu, Chodorow, M, Braden-Harder, L and Harris, M D (1998b) Automated scoring using a hybrid feature identification technique, in Proceedings of the Annual Meeting of the Association of Computational Linguistics, August 1998, Montreal, Canada.

Burstein, J, Leacock, C and Swartz, R (2001) Automated evaluation of essays and short answers, paper presented at 5th International Computer Assisted Assessment Conference, Loughborough University.

Callear, D, Jerrams-Smith, J, and Soh,V (2001) CAA of short non-MCQ answers, in Proceedings of the 5th International CAA conference, Loughborough. Retrieved from: http://www.lboro.ac.uk/service/ltd/flicaa/conf2001/pdfs/k3.pdf.

DeLoughry, T J (1995) Duke professor pushes concept of grading essays by computer, *Chronicle of Higher Education*, ppA24–25.

Foltz, P W (1996) Latent semantic analysis for text-based research, *Behavioral Research Methods, Instruments and Computers*, 28, 197–202.

Foltz, P W, Britt, M A and Perfetti, C A (1996) Reasoning from multiple texts: An automatic analysis of readers' situation models, in Proceedings of the 18th Annual Cognitive Science Conference, 110–115. Hillsdale, NJ: Lawrence Erlbaum Associates.

Foltz, P W, Laham, D and Landauer, T K (2003) Automated essay scoring: Applications to Educational technology. http://www-psych.nmsu.edu/~pfoltz/reprints/Edmedia99.html. Reprint.

Graesser, A C, Wiemer-Hastings, P, Wiemer-Hastings, K, Harter, D (2000) Person N and the Tutoring Research Group. Using latent semantic analysis to evaluate the contributions of students in AutoTutor, *Interactive Learning Environments*, 8(2), 87–109.

Kanejiya, D, Kumar, A and Prasad, S (2003) Automatic evaluation of students' answers using syntactically enhanced LSA, in *Building Educational Applications Using Natural Language Processing, Proceedings of the HLT-NAACL 2003 Workshop*, 53–60, Association of Computational Linguistics.

Larkey, L S (1998) Automatic essay grading using text categorization techniques, in Proceedings of the Twenty First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 90–95.

Marcu, D (2000) *The Theory and Practice of Discourse Parsing and Summarization*, The MIT Press.

Maron, M E (1961) Automatic indexing: An experimental enquiry, *Journal of the Association for Computing Machinery*, 8, 404–417.

Mitchell, J S (1998) Commentary: SAT don't get you in, *Education Week on the Web*. Retrieved from: http://www.edweek.org/ew/current/3mitch.h17

Mitchell, T (1997) *Machine Learning*, Maidenhead: McGraw-Hill.

Mitchell, T, Russell, T, Broomhead, P and Aldridge, N (2002) Towards robust computerised marking of free-text responses, in Proceedings of 6th International Computer Aided Assessment Conference, Loughborough.

Page, E B (1966) The imminence of grading essays by computer, *Phi Delta Kappan*, January, 238–243.

—(1994) Computer grading of student prose, using modern concepts and software, *Journal of Experimental Education*, 62, 127–142.

Perfetti, C (1998) Teachers of Tomorrow? Retrieved from: http://www.wired.com/news/technology/0,1282,16009,00.html

Qian, D D and Harley, B (1998) The role of depth knowledge in adult ESL learners' comprehension of academic texts, paper presented at the Conference of the Canadian Association for Applied Linguistics on Trends in Second language Teaching and Learning, Ottawa, Canada.

Rosé, C P, Roque, A, Bhembe, D and VanLehn, K (2003) A hybrid text classification approach for analysis of student essays, in *Building Educational Applications Using Natural Language Processing*, 68–75.

Salton, G (1989) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Reading, MA: Addison-Wesley.

Schwartz, A E (1998) Graded by machine, *Washington Post*, p.107.

Sleator, D K and Temperley, D (1991) Parsing English with a link grammar, Technical Report. October CMU-CS-91–196.

—(1993) Parsing with a link grammar, in Proceedings of Third International Workshop on Parsing Technologies, Tilburg, the Netherlands, August 10–13.

Wiemar-Hastings, P and Zipitria, I (2001) Rules for syntax, vectors for semantics, in Proceedings of 23rd Annual Conference of the Cognitive Science Society, Mahwah, N J: Laurence Erlbaum Associates.

Wolfe, M, Schreiner, M E, Rehder, B, Laham, D, Foltz, P W, Kintsch, W and Landauer, T K (1998) Learning from text: Matching readers and texts by latent semantic analysis, *Discourse Processes*, 25(2 and 3), 309–336.

Wresch, W (1993) The imminence of grading essays by computer – 25 years later, *Computers and Composition*, 10, 46–58.

Studies in Language Testing – Volume 18

Volume 18 in the *Studies in Language Testing* series is entitled *European Language Testing in a Global Context*. It contains a small subset of the conference papers presented at the first ALTE conference held in July 2001 in Barcelona, Spain. 2001 was designated as the European Year of Languages and the ALTE conference was held in celebration of this event.

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

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

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

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

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

The second part of the volume presents a number of research studies. Anne Lazaraton focuses on the use of qualitative research methods in the development and validation of language tests. Lazaraton is a pioneer of qualitative research in language testing and her involvement dates back to the late eighties when such approaches were not yet widely used in the field. It is in part due to her efforts that researchers are now more willing to embrace approaches that can provide access to the rich and deep data of qualitative research. Readers are encouraged to look at an earlier volume in the *Studies in Language Testing* series written by Lazaraton (*A Qualitative Approach to the Validation of Oral Language Tests* Volume 14).

Vivien Berry and Jo Lewkowicz focus on the important issue of compulsory language assessment for graduating students in Hong Kong. Their paper considers alternatives to using a language test alone for this purpose and looks at the applicability of variations on the portfolio concept. Jim Purpura's work on the validation of questionnaires, which addresses the interaction of personal factors and second language test performance, represents an interesting and challenging dimension of validation in language testing. An earlier volume by Purpura in the *Studies in Language Testing* series (*Learner strategy use and performance on language tests:*

A structural equation modelling approach Volume 8) looks in more depth at the development of questionnaires to determine personal factors and a methodology that can be used to investigate their interactions with test performance.

Annie Brown reports on an exploratory study investigating the salience of legibility as a factor in rater judgements. This is particularly relevant as we move towards greater use of computers in language testing. Such a move is, of course, fraught with issues, not least of which is the one of legibility that Brown addresses here. In the context of IELTS and other Cambridge ESOL tests, such research is currently being conducted in Cambridge and has been reported in *Research Notes*.

Barry O'Sullivan's paper attempts to model the factors affecting oral test performance, an area of particular significance in large-scale assessment. The paper is part of ongoing research commissioned by Cambridge ESOL and a collection of research studies into the dimensions of oral assessment will be published in due course.

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

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

Conference Reports

We report below on a range of conferences attended by Cambridge ESOL staff and consultants: Reading Worlds (Lynda Taylor and Diana Fried-Booth), Writing Developments in Higher Education and BALEAP PIM on Phraseology and Metaphor in EAP (Stuart Shaw) and Lynda Taylor reports on a Young Learners Symposium that Cambridge ESOL hosted recently.

'Reading Worlds' – British Council Conference on the teaching of literature

The 19th Oxford Conference on the teaching of literature – organised by the British Council Seminars programme – was entitled 'Reading Worlds' and was held at Corpus Christi College in Oxford from 2nd to 8th April 2004. Thirty-nine participants attended the seminar from the following countries: Ireland, Denmark, Sweden, Norway, Spain, Portugal, Germany, Switzerland, Austria, Slovenia, Slovakia, Romania, Macedonia, Russia, Kyrgyzstan, Malta, Cyprus, Morocco, Lebanon, India, Bangladesh, Australia, USA, Venezuela, and Brazil. Most were school teachers or university lecturers, and all had a strong professional and personal interest in language and literature.

The Oxford Conference has influenced discussion and debate on literature teaching for nearly 20 years. Participants have opportunities to meet and hear some of the most eminent academics, critics and writers in the UK, as well as to take part in discussions about a range of pedagogic and literary trends and issues. In bringing together the insights and experiences of participants and speakers in a strongly international context, the Oxford conference offers an unrivalled event. The 2004 programme featured two interrelated professional tracks, each exploring two important facets in literature study and teaching. The first looked at the 'politics' of contemporary canon-building and how reputations are built. The second examined assessing the value of literature teaching. The importance of readers, and reading as a creative activity, linked both tracks.

Representatives from Cambridge ESOL were invited to contribute to the discussion sessions focusing on assessment and evaluation which took place on Wednesday 7 April.

In the morning Dr Lynda Taylor took part in the Assessment Panel chaired by Alan Pulverness. Other panel members were: Julia Bell (novelist and lecturer in Creative Writing at Birkbeck College, University of London); Isaac Quist (Curriculum Area Manager for the International Baccalaureate Organisation); Noel Cassidy (secondary English teacher, and a senior examiner for A/AS level English for both OCR and CIE). A range of questions on assessment issues had been gathered from participants in advance and these were submitted to the panel for comment and discussion; questions related to: the purpose of assessment; the suitability of different testing/assessment formats; defining the construct of reading; alternative assessment approaches (e.g. portfolios); issues of validity, reliability, practicality and impact; the tension between assessing language and content.

In the afternoon Diana Fried-Booth, an experienced ELT consultant and Chair of the CPE Paper 2 set texts, presented a session as part of the Exam Board Showcase. Her presentation dealt with the rationale for including set texts in FCE and CPE which are a long-established feature of the Cambridge ESOL examinations. She referred to the tension in the assessment of a candidate's response to literature in an examination which focuses on assessing a candidate's ability to communicate. Reference was made to the General and Task Specific Markschemes in looking at the criteria which underpin assessment. The factors which have to be taken into account in selecting appropriate texts for an international candidature were also considered. A brief historical overview from 1914 onwards was provided, in order to reflect the shift that has

taken place over the years in the choice of texts, particularly in the last eight years. The presentation continued with a review of the advantages to be gained by language teachers and students in studying a set text, and the ways in which language teachers can exploit the potential of literature. In conclusion the presentation looked at possible choices of different genres for the future and invited teachers to send in suggestions they might have for texts to the Cambridge ESOL website or to the appropriate Subject Officer.

In a future issue of *Research Notes* we hope to discuss at greater length the role of the literary set texts option in the Cambridge examinations.

WDHE (Writing Developments in Higher Education) 2004: Valuing and Evaluating Writing in Higher Education

The WDHE initiative was launched in September 1994 by the University of Northumbria in response to widely expressed concerns about the standards of students' written communication skills. Previous conferences have brought together practitioners and researchers from a wide range of countries, institutions and research areas to discuss the forms that writing developments can take within Higher Education.

The WDHE 2004 conference at Sheffield Hallam University from 11–12 May focused on how (and how much) students' writing is valued, assessed, and evaluated, as well as celebrating the 10th conference and a decade of significant development.

Specific themes and issues included:

- Writing for employability – how the writing skills students need for their courses can be developed in a way which encourages their transfer to the workplace
- Academic writing – how it can be valued and how space can be created to accommodate writing development in an already crowded curriculum
- Assessment strategies, methods and criteria – how Higher Education assessment practices can support students' valuing of their own writing and its development
- International perspectives – how writing is valued and evaluated in different national contexts
- New technologies – how they can be used to help students develop their academic writing (and how they can also help staff support students)
- Student and staff diversity – how the needs of students and staff can be met
- Staff development and university structures – how staff across different disciplines can help students develop their skills and how this can be facilitated.

Academic writing is increasingly the focus of debates in many countries and the questions being raised are of importance to Cambridge ESOL – particularly with regard to the academic writing dimension of IELTS. Questions the conference presentations attempted to answer included:

- Are the issues that are encountered similar in different contexts? How do they differ?
- What kinds of questions are being raised in different contexts and what kinds of answers – theoretical, pedagogical, institutional – are emerging?
- Given different historical, linguistic and cultural contexts, to what extent can research findings and pedagogical programmes developed in one context be usefully adopted to others?

The conference provided an opportunity for Cambridge ESOL to contribute to the on-going debate on the assessment of academic writing skills together with an overview of current EAP teaching practice and pedagogy. Additionally this event gave us the chance to correct any misconceptions about our range of examinations, specifically the IELTS Academic Writing module. It also enabled us to evaluate how well Cambridge ESOL Writing Tests reflect current developments in EAP, and specifically developments in teaching English to the target candidature, namely, candidates intending to study in an English-speaking Higher Education environment.

BALEAP Professional Issues Meeting: Phraseology and Metaphor in EAP

A BALEAP (*British Association of Lecturers in English for Academic Purposes*) Professional Issues Meeting was held at the University of Birmingham on 19 June 2004. The one-day event focused on the topic of Phraseology and Metaphor in EAP.

David Oakey and Jeannette Littlemore (English for International Students Unit, University of Birmingham) opened the meeting with an overview of these two areas of language which cause difficulty for university learners of English. David and Jeannette considered different conceptual approaches to identifying phraseology and metaphor, methodological approaches to observing and identifying phraseological and metaphorical realisations generally and then in specific terms, the use of phraseology and metaphor by academic discourse communities, and pedagogical applications of research whereby students can be helped to understand and use phraseology and metaphor in EAP contexts. Areas of overlap between phraseology and metaphor raised issues that were addressed by other speakers during the day.

In his presentation – Consolidating collocational competence – Peter Howarth (University of Leeds) attempted a survey of current methods for assisting learners to remedy their collocational errors. Collocational, or phraseological, competence, he argued, is perhaps one of the highest levels of linguistic proficiency that learners can attain. While, he contended, for many purposes absolute native-speaker naturalness is not always required (or even possible) there are some learners who aim at high levels of achievement in this area for occupational purposes and this is an area receiving great attention in EFL and EAP coursebooks and in examinations such as Cambridge ESOL's Certificate of Proficiency in English. The view presented here was that collocations themselves cannot be taught; instead, teachers should focus on raising awareness of how the phenomenon of phraseology operates in the language, especially its connection with formal style, and should encourage students to make use of the range of reference materials available.

John Skelton and Andrew Shanks (Department of Primary Care and General Practice, University of Birmingham) considered metaphoric expressions in two health-care settings. They focused on differences between health professional and service-users in their use of metaphoric expressions in two separate contexts: doctor-patient consultations in general practice and focus groups consisting of mental health service users, community psychiatric nurses and general practitioners. For both studies, implications for training were explored.

Almut Koester (University of Birmingham) presented on teaching idioms in spoken academic discourse, arguing that idioms play an important role in this spoken genre and often present a particular problem for non-native speakers; not only because they involve metaphors and multi-word units, but also because of the indirect and implied meanings and functions they convey. Some preliminary attempts to teach idioms from a functional perspective to international students were reported and evaluated.

Julia Lockheart (Goldsmiths College, University of London) and Harriet Edwards (Royal College of Art) presented a series of visuals, both diagrammatic and pictorial in nature, that reflect ways in which tutors and students employ metaphors in their teaching and learning practices around writing in the arts/humanities context.

Willie Henderson (University of Birmingham) explored how the work of applied linguists reveals the significance of metaphor in Economics and how economic methodologists also look at the significance of metaphor in Economics. In his presentation he attempted to ascertain what, if anything, can they learn from each other.

Svenja Adolphs (University of Nottingham) looked at the association between high and low levels of social-cultural integration amongst international students at English-medium universities and the acquisition of recurrent lexical clusters. Whilst it is widely held that exposure to language plays a significant part in the acquisition process especially with regard to the acquisition of individual lexical items, grammatical structures, and discourse competencies the influence that exposure has on the acquisition of formulaic sequences in language use has been less well demonstrated. Svenja's presentation explored how social/cultural integration (and the related exposure) affects the acquisition of formulaic sequences.

Philip King (University of Birmingham) gave an interesting seminar entitled Cross-cultural Wiring in which he assessed writing by a Greek student working at MA level. The writing showed a number of problems and exhibited unusual patterns of lexical phrases, in some cases under the influence of the student's first language. The presentation explored the profiles of some of these expressions in both languages and aimed to show how despite superficial similarities between some expressions in both languages, the collocational boundaries are not coterminous.

The event closed with a presentation by Frank Boers (Erasmushogeschool Brussels, Belgium) entiltled Phrase-wise English: some rhyme and reason. Mastery of semi-fixed lexical phrases (idioms, collocations, etc.) is believed to facilitate L2 fluency and helping learners notice useful expressions in the authentic L2 language they are exposed to adds a valuable dimension to L2 language instruction. In his presentation Frank reported on a series of experiments whose aim was to assess the efficiency of mnemonic strategies that could make the task of memorising phrases a little easier for students. In addition, he also presented evidence of beneficial effects of the given strategies with regard to comprehension, register awareness and cultural awareness.

This event illustrated the range of research being undertaken on metaphor and phraseology in EAP settings and provided Cambridge ESOL with insights into how their exams are viewed by the EAP community in relation to these areas. Incorporating collocational knowledge into teaching and testing, particularly at higher proficiency levels, is clearly a challenging enterprise. We continue our research into describing written and spoken proficiency using corpora in order to explore this area to ensure that all Cambridge exams remain relevant to the needs of test takers and users and form the cutting edge of language testing design and implementation.

Cambridge ESOL's Young Learners Symposium – May 2004

It is over 10 years since Cambridge ESOL began work on a suite of language proficiency tests for younger learners of English. The Young Learners English (YLE) Tests at three levels – *Starters, Movers* and *Flyers* – are designed for 7 to 12 year-olds and aim to achieve the following:

- sample relevant and meaningful language use
- measure accurately and fairly
- promote effective learning and teaching
- offer a positive first impression and experience of international tests.

The original design and development of our YLE tests involved an extensive review of the available research literature in a number of relevant fields, as well as consultation with a wide range of stakeholders, including members of the academic, pedagogic and publishing communities. Following successful trialling in various parts of the world, the tests became operational in 1997. Interest in the YLE tests has grown steadily since then; by the end of 2003 annual candidature had reached almost 300,000, with large numbers of candidates in China, Spain, Argentina and Italy.

Since 1997 the operational test performance of the YLE Tests has been routinely monitored within Cambridge ESOL; an annual *YLE Examination Report* is also produced for external publication to provide test users with an overview of test performance during a given year. These ongoing monitoring and evaluation activities are an essential part of the validation process for any test; they provide evidence that a test is functioning as intended, and they can help to identify features of the test which may need to be reviewed and revised at a later stage. Regular consultation with test users is a priority and reflects the importance Cambridge ESOL attaches to evaluating the washback and impact of its tests, especially those for young learners. In 2002 – five years after their introduction – we began a formal review process with a view to making some minor changes to the tests from 2007. Findings from some of the consultation exercises and the internal studies of test performance

carried out under this review have been reported in previous issues of *Research Notes* (see, in particular, issues 7 and 15).

Cambridge ESOL has a long tradition of engaging with members of the various 'stakeholder' communities who use or have an interest in our tests, especially those who are practitioners in the world of language teaching/learning and those who work in an academic research context. Regular consultation with specialists from these groups is an important element in our approach to developing and revising all of our tests (see, for example, the published account of the CPE Revision Project, 2003). At a meeting of the internal YLE Steering Group in 2002, we discussed how we might engage more effectively with the growing community of individuals working throughout the world in the field of young learner teaching, learning and assessment. It was the late Dr Peter Hargreaves, then Chief Executive of our organisation, who first suggested that we organise an event to which we could invite teachers, trainers, academic researchers, materials writers and language testers with an active interest in young learners of English and other foreign languages. By bringing together in Cambridge a group of specialists with professional and personal interest in young language learners, he believed we might create a context where knowledge and experience could be shared, current issues of interest/concern could be discussed, and new professional relationships established.

The conceptualisation of such an event took time to evolve and in the end we decided to plan an invited symposium on the general theme of "Issues in the teaching, learning and assessment of L2 young learners". Rather than organise a traditional, conference-style event, we chose a slightly more experimental approach: invited participants were commissioned to submit a paper relating to one of four main themes; the submitted papers were then compiled into a delegate volume which participants were asked to read prior to the symposium. We hoped that this approach would encourage a stimulating debate of some key issues and concerns, and would allow plenty of time for interaction during the actual symposium meeting. Finally, after many months of planning, Cambridge ESOL's YL Symposium took place on 28/29 May 2004 at Downing College in Cambridge University. Around 20 external delegates were able to attend: some represent key universities and other organisations in the UK; many live and work in Europe (France, Spain, Portugal, Hungary, Cyprus, Greece); and some live and/or have experience of working in other parts of the world (e.g. Canada, India, China). The external delegates were joined by a similar number of Cambridge ESOL staff, many of whom work directly on our YLE Tests.

To open the symposium, Lynne Cameron (Professor of Applied Linguistics in the School of Education at the University of Leeds) delivered an introductory plenary talk on the theme of 'The learning child in the language classroom'. In a fascinating and stimulating paper she offered a new metaphor for thinking about young learner issues by drawing on 'complex systems theory' – a conceptual framework which several other academic disciplines have found useful (e.g. biological sciences, economics, management). This set the scene for four 90-minute discussion sessions during the symposium which attempted to explore the cognitive, social, pedagogic and assessment dimensions of language teaching and learning among children. Each session was chaired by a member of the Cambridge ESOL staff and was structured around a set of focus questions derived from the papers in the delegate volume. Throughout our discussions we found ourselves returning to some of the fundamental concepts of complex systems theory – dynamism, variation, interconnectedness and creativity – as they apply in the context of language learning and teaching among children. To close the symposium, Melanie Williams – an experienced teacher trainer and consultant specialising in young learner education – skilfully drew together the various threads of our discussions to reflect the key elements and identify possible directions for the future.

Although somewhat experimental in its content and format, the YL Symposium proved stimulating and fruitful from a number of different perspectives. First of all, it offered a rare opportunity for members of the YL academic, pedagogic and assessment community to meet one another and spend time together; the sharing of knowledge/experience and the building of professional relationships between individuals and groups went on not only during the plenaries and discussion sessions, but also over the coffee/lunch breaks and during the formal dinner on the Friday evening. Even though professional colleagues meet on the international conference circuit, there is rarely much time to sit down and discuss issues of mutual interest in any depth. Secondly, the event offered a valuable opportunity for the community to share and debate matters of policy and practice in the teaching, learning and assessment of young language learners. It soon became clear that absolute principles and definitive conclusions were hard to reach, largely because of the many varied and complex environments around the world in which teaching/learning takes place. However, this simply confirms the need for more research to be undertaken in the field and for a continuing professional debate about what constitutes best practice within specific contexts. Finally, from Cambridge ESOL's perspective we welcomed the chance to consult with some key stakeholders in the YL field, to explain our approach to young learner assessment, and to hear the issues which concern them. The views expressed will be invaluable in informing our own thinking and activity as test providers.

As a result of the symposium, plans are already in hand for increased collaboration between Cambridge ESOL and individuals/institutions, both in the context of the YLE Review and more broadly. One key objective over the next few months will be a project to develop the draft papers prepared for the symposium into an edited volume in the Studies in Language Testing series. We hope that this will be a valuable contribution to the field, not just for academics and researchers but also for teachers, teacher-trainers and others working in the area of YL teaching, learning and assessment.

Other News

Launch of English UK

The two leading associations in UK English language teaching – ARELS (Association of Recognised English Language services) and BASELT (British Association of State English Language Teaching) – have joined forces to form a new association called English UK. This is an important development in language teaching and creates a major new force to promote the UK as an English language study destination for students worldwide.

The launch of English UK took place on 12 May 2004 in London and was followed by a Parliamentary Reception at the Houses of Parliament. The day featured a number of interesting presentations by experts in the field, including representatives from English UK and the Department for Education and Skills. Dr Michael Milanovic, Chief Executive of Cambridge ESOL, spoke about innovations in language testing and represented Cambridge ESOL at the Parliamentary Reception. He is pictured below with Simon Lebus, Chief Executive of the UCLES Group and Charles Clarke, UK Secretary of State for Education.

From left: Michael Milanovic, Chief Executive of Cambridge ESOL, and Simon Lebus, Chief Executive of the UCLES Group, meet Charles Clarke, UK Secretary of State for Education, at the Parliamentary reception.

As Cambridge ESOL shares English UK's commitment to ensuring quality in English language teaching, it was proud to sponsor the launch and reception.

To find out more visit the English UK website: www.englishuk.com

New Certificates in ESOL Skills for Life

Skills for Life is the UK's national strategy for improving adult literacy, numeracy and ESOL skills. From September 2004, all courses and qualifications for adult learners of English in the UK state sector must be based on the Adult ESOL Core Curriculum.

In response to these requirements, Cambridge ESOL is offering three new Certificates in ESOL Skills for Life, which provide flexible assessment of ESOL learners in the UK, whilst minimising the workload of teachers and administrators. They reflect the use of English in everyday life within the UK and are fully based on the Core Curriculum.

The new Skills for Life Certificates are:

- available at three levels – Entry 1, Entry 2 and Entry 3
- externally assessed by trained and experienced examiners
- available on demand to suit the learner and the learning provider
- accredited by the QCA and funded by the Learning and Skills Council
- available from March 2005 to support programmes commencing September 2004.

Recognising that the individual's language skills can vary quite significantly in their development, the new Certificates allow learners to enter for different modes (Reading, Writing, and combined Speaking and Listening) at different levels. These modes reflect the requirements of the Core Curriculum and will enable candidates to build up a portfolio of achievement over time.

At each level, learners who achieve a pass will be awarded certificates from Cambridge ESOL. After Entry 3, candidates can progress onto other qualifications. Cambridge ESOL offer qualifications at Level 1 (FCE, CELS Vantage, BEC Vantage), Level 2 (CAE, CELS Higher, BEC Higher) and Level 3 (CPE).

The Certificates are designed for adult (16+) speakers whose home language is not English, who are living or trying to settle in the UK. It is expected that the learners' educational and employment backgrounds will be diverse, as well as their aspirations, literacy levels and language learning skills. This diversity will be reflected in the range of material selected for use in these tests and in the task types which candidates will need to complete.

Further information on the tests, along with classroom materials and notes on the preparation of candidates, will be available on the Cambridge ESOL Teaching Resources website from Autumn 2004.

Please visit the Skills for Life website: www.CambridgeESOL.org/SFL/

Teaching Knowledge Test (TKT)

Cambridge ESOL is developing a new test for teachers of English which focuses on the core teaching knowledge needed by teachers of primary, secondary or adult learners, anywhere in the world.

TKT is suitable for teachers with a level of English at Level B1 of the Common European Framework (PET level) or higher. The test will be available in 2005 and is currently undergoing extensive trialling throughout the world. Look out for articles on TKT in future issues of *Research Notes*.

Changes to Cambridge ESOL's provision for candidates with dyslexia

An earlier issue of *Research Notes* (11 – February 2003) reported on the special arrangement provisions for a range of candidates, including dyslexic candidates. For Upper Main Suite Writing papers (FCE, CAE and CPE), our provisions for candidates with dyslexia currently include the option of separate marking. As part of our commitment to keep all aspects of our Special Arrangements policy and practice under constant and careful review, Cambridge ESOL has conduced research into the appropriacy of this provision.

The number of theoretical and empirical studies in the area of dyslexia has grown considerably in recent years and Cambridge ESOL's own studies have revealed that 'separate marking' of writing is not an appropriate special arrangement in language-focused assessment (even though it is common practice in content-focused assessment).

For this reason, from December 2004, Cambridge ESOL will no longer offer separate marking as a provision for candidates with dyslexia. This change in policy is consistent with expert advice in the field and with the policy/practice of other assessment providers.

The 'extra-time' provision is generally acknowledged to be the most appropriate provision for candidates with dyslexia and this will continue, as will the other special arrangements which are available.

2003 Grade Statistics now available

Have you ever wondered how candidates score worldwide on their Cambridge ESOL examinations? On our website you can find annual grade statistics for a wide range of examinations showing the average percentages in each grade for all sessions by country (please note these statistics do not include absent, partial absent or ungraded candidates).

The latest statistics, for the 2003 calendar year, can be found at: www.cambridgeesol.org/stats/2003/index.cfm

Call for proposals for ALTE 2nd International Conference, Berlin, May 2005

The Association of Language Testers in Europe (ALTE) is holding its second major international conference from 19–21 May 2005 at the DBB Forum, Berlin, Germany.

The aim of the conference is to provide a forum for the discussion of issues associated with:

- 'Language Assessment in a Multilingual Context: Attaining Standards, Sustaining Diversity'
- The relationship of linguistic diversity to the testing of languages in Europe
- The interaction of the above areas with developments in diversity and testing in a global context.

This event is open to the public and ALTE invites papers on the following themes and topics of discussion:

Quality and Diversity

- Quality of examinations and Codes of Practice
- Setting standards in testing: minimum standards
- Research on impact
- Testing language for specific purposes
- State certification projects
- The testing of young learners: children of primary and middle school age
- Issues related to the testing of less widely spoken languages
- IT/Distance Learning

Ethics and Diversity

- Language and citizenship evaluation
- Language testing and immigration

Transparency and Diversity

- Relating examinations to the Common European Framework of Reference
- Use of the Council of Europe Pilot Manual for Relating Examinations to the Common European Framework

If you would like to consider submitting a paper proposal, please visit the ALTE website (www.alte.org) where you will find the submission form and accompanying guidelines. All paper proposals must be submitted electronically by **Thursday 30 September 2004** via the ALTE website. The acceptance of individual proposals will be announced in January 2005.