CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

100 CAMBRIDGE ENGLISH
CENTENARY 1913–2013

Research Notes

Issue 52

May 2013

ISSN 1756-509X

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Research Notes

Issue 52 / May 2013
A quarterly publication reporting on research, test development and validation

Printed in the United Kingdom by Océ (UK) Ltd.

Research Notes

Contents

Editorial notes

Welcome to issue 52 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge English Language Assessment.

This issue features articles on the first European Survey on Language Competences (ESLC) and from the second round of the Cambridge English Funded Research Programme.

The issue starts with Dr Neil Jones, the Director of the ESLC, providing an overview of the project, summarising the main findings and describing the implications of these findings for educational policy makers and Cambridge English Language Assessment. Jones highlights the importance of this survey and how it confirms the widely held views that language learning is successful when language is used for communicative purposes. Martin Robinson, who was the Manager of Test Development for the survey, then describes the language testing framework that was used as the basis to develop the ESLC survey instruments. His article details the challenges that had to be overcome when developing comparable language tests in five languages. The next two articles describe how country-specific ESLC data is being used for secondary research and to inform national policy decisions. First, Magdalena Szpotowicz, the ESLC national research coordinator for Poland, discusses the Polish experience while her counterpart in Croatia, Jasminka Buljan Culej, follows with the Croatian experience. Karen Ashton, the ESLC Project Manager, reflects on the lessons learned from this first survey and offers recommendations for the second survey. Stephen McKenna then reports on the public reaction to the ESLC.

The issue then moves on to highlight studies undertaken in 2011 as part of the Cambridge English Funded Research Programme (Round 2). The first is by Anthony Green, Hanan Khalifa and Cyril J Weir, which explores the features that distinguish reading texts at three proficiency levels using *Coh-Metrix*. The second, by Okim Kang, investigates criterial features that can be used to distinguish aspects of spoken language at different CEFR levels. The findings from these research studies can inform both test development and classroom instruction.

The European Survey on Language Competences and its significance for Cambridge English Language Assessment

NEIL JONES RESEARCH AND VALIDATION GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Outline of the European Survey on Language Competences

The first European Survey on Language Competences (ESLC) published its findings in June 2012. The survey was set up by the European Commission to measure levels of achievement in foreign languages in European secondary schools, and also to explore the relationship between language proficiency and contextual factors such as onset of learning, language learning environment, use of information and communication technologies (ICT), study abroad, teacher training and teaching methods, with the aim of providing insights into good practice. The results of the survey will also enable the European Commission to establish a European language competence indicator to measure progress towards the 2002 Barcelona European Council Conclusions, which called for 'action to improve the mastery of basic skills, in particular by teaching at least two foreign languages from a very early age' (European Commission 2005).

SurveyLang, the multinational consortium which won a competitive tender to conduct the survey, had the following members: Cambridge English Language Assessment (project management, English language tests, coordination of language test development); Centre international d'études pédagogiques (CIEP) (French language tests); Goethe-Institut (German language tests); Università per Stranieri di Perugia (Italian Language tests); Universidad de Salamanca/Instituto Cervantes (Spanish language tests); Gallup Europe (sampling, testing tool development, translation); and the National Institute for Educational Measurement (Cito) (questionnaire design, analysis). Development work started in 2008 and the main survey was administered in 2011.

Over 53,000 students across 14 European countries took part in the survey. Belgium's three linguistic communities participated separately to give a total of 16 educational systems. This is a reasonable number for the first administration of a new survey (initiated in difficult economic circumstances) and certainly sufficient to provide interpretable results.

Some key features of the survey were determined by the Terms of Reference provided by the European Commission. The languages included were the five most widely taught in Europe: English, French, German, Italian and Spanish. Within each educational system the two most-taught of these were to be tested. The target student group was the last year of lower secondary education (age 13–15, depending on country); however, the second year of upper secondary was chosen (age 15–16) where a language was not taught earlier than this. The survey covered the three skills of listening, reading and writing, speaking being excluded in this first round due to concerns as to the practicability of testing it.

Language test results

For a fuller presentation of the results see the ESLC Final Report (European Commission 2012a). Figure 1 and Figure 2 show the percentage of students achieving each Common European Framework of Reference (CEFR) level from Pre-A1 (i.e. failing to achieve A1) up to B2 (the highest level tested in the survey) for each country. In these figures the results are summarised across the three tested skills by taking an average of the percentage achieving each level in each skill.

The countries are shown ordered from the highest performing (i.e. having more students at higher CEFR levels and fewer at low levels) to the lowest performing. This has the advantage of clear presentation, but the disadvantage that it suggests a simple 'league table' approach to evaluation. In fact, as we discuss further below, contexts of language learning differ so greatly across countries and languages that to understand the situation in a given country requires a much more qualitative and differentiated approach to evaluation.

None the less, the bare language test results tell a story: there is clearly a very wide range of achievement across countries and education systems. Figure 1 presents results in first foreign language, which is English (EN) for all countries except England itself, and the Flemish and German communities within Belgium, for whom it is French (FR). Figure 1 shows that Sweden is the highest performing country, with 57% of students achieving CEFR Level B2. England is the lowest performing country, with 30% of students failing to achieve even CEFR A1, and only 2% achieving B2.

Figure 2 presents the picture for second foreign language. It can be seen that German (DE) is the second foreign language in eight education systems, French (FR) in three, Spanish (ES) in two and Italian (IT) in just one. Note that 'first' and 'second' here relate to the five tested languages.

The picture for second foreign language shows the same wide range of achievement as first foreign language. Achievements in CEFR terms are somewhat lower, reflecting in part the generally much shorter duration of learning.

For both first and second languages the number of students achieving no higher than A1, or not even achieving that, is high in many countries.

Questionnaire results

The questionnaires are organised around a number of language learning policy issues identified as being of interest to the European Commission. The ESLC presents the questionnaire findings in two ways: as simple tabulations of the data by country – for example, showing the frequency of

Figure 1: First foreign language: Percentage of pupils at each level by educational system using global average of the three skills

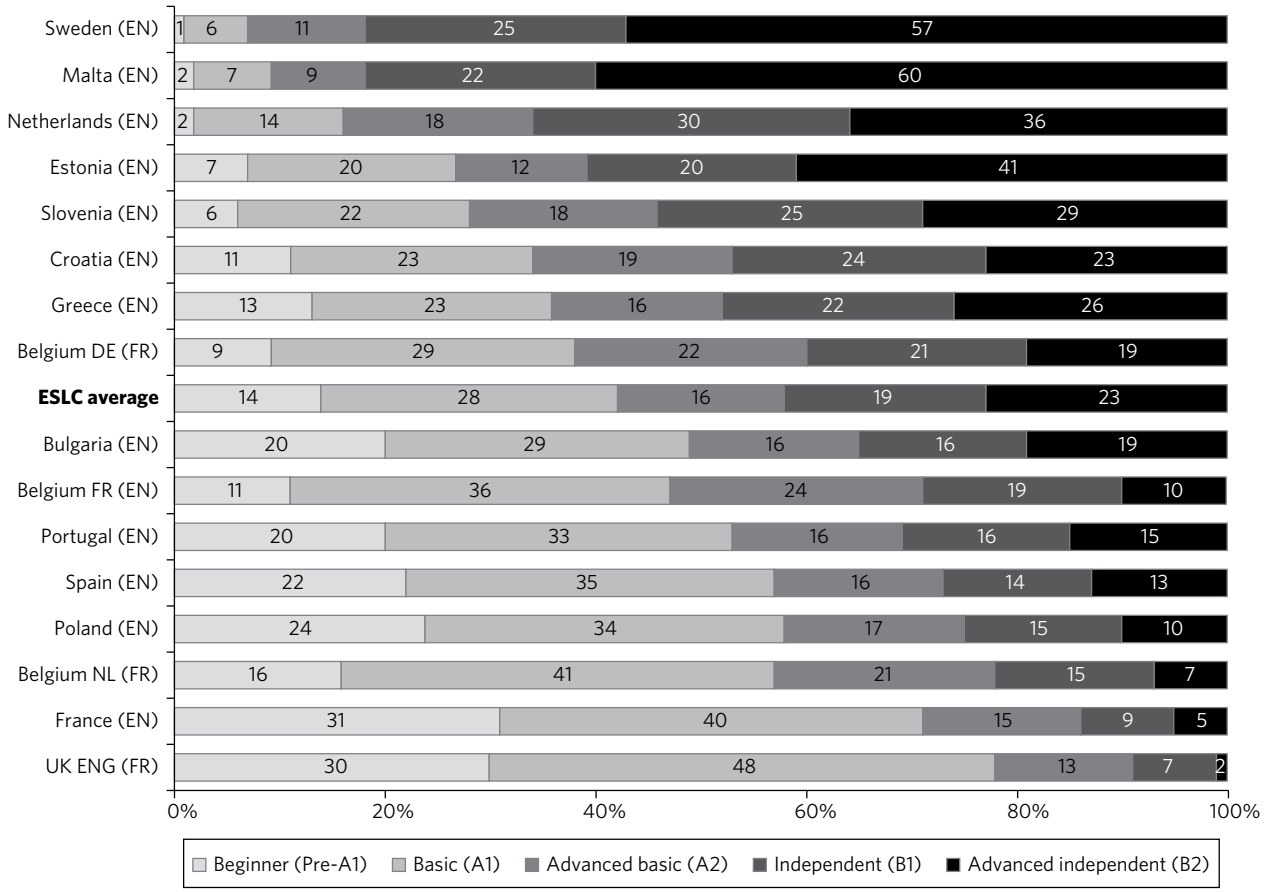
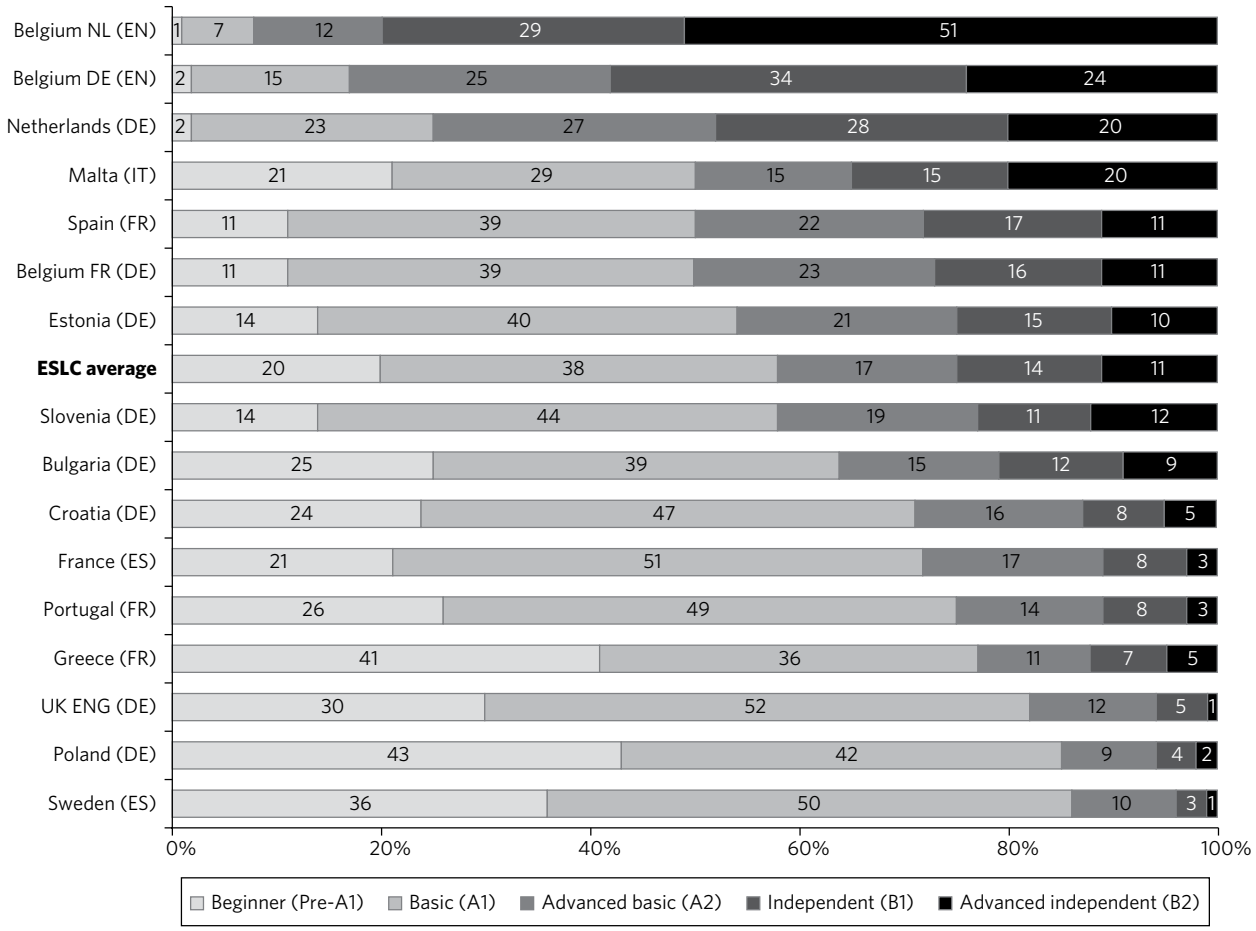


Figure 2: Second foreign language: Percentage of pupils at each level by educational system using global average of the three skills



holiday trips abroad as reported by students – then regression analysis is used to explore the relationship between such data and the outcomes of the language tests. This might show for example that there is a significant positive relationship between taking holiday trips abroad and achievement in learning the language of that country.

For many of the policy issues addressed it was possible to show significant relationships with language test outcomes. For others it was not possible, simply because countries did not differ much regarding the specific issue at question. Where there is little difference, perhaps because all countries stand equally high or low on some index, then the regression will not detect an effect, even if the effect exists. Thus, for example, it was not possible to show a positive impact of using content and language integrated learning (CLIL), because it is generally still rather rarely adopted; but this does not prove that CLIL is not an effective approach to language learning.

Another necessary limitation on how we may interpret regression findings is that we should not treat a significant relationship as a proof of causation. For example, the survey found a significant relationship between the number of languages studied and ability in the tested language. But we cannot infer that studying more languages makes you a good language learner; it might be that people study more languages because they are good language learners.

The contextual information collected through the questionnaires focuses on those contextual factors which can be modified through targeted educational policies, such as the age at which foreign language education starts, or the training of teachers. The ESLC maps out differences within and between educational systems regarding three broad policy areas, and evaluates which of these relate to differences in language proficiency. Other factors which are largely beyond the control of policy such as general demographic, social, economic and linguistic contexts are not explicitly discussed in the final report, although data on socio-economic status is collected and is available for analysis by educational systems. The following brief summary of the questionnaire findings is adapted from the ESLC executive summary (European Commission 2012b).

An early start to language learning

Generally pupils report a rather early start to foreign language learning (before or during primary education) and most commonly they learn two foreign languages. However, considerable differences are still found across educational systems in the exact onset of foreign language learning, the current teaching time and the number of languages offered and learned.

The results of the ESLC show that an earlier onset is related to higher proficiency in the foreign language tested, as is learning a larger number of foreign languages and of ancient languages.

A language-friendly living and learning environment

Policy also aspires to create a language-friendly living and learning environment, where different languages are heard and seen, where speakers of all languages feel welcome and language learning is encouraged. Clear differences between educational systems are seen in the informal language learning opportunities available to pupils (such as pupils' perception of their parents' knowledge of the foreign language tested, individual trips abroad, the use of dubbing or subtitles in the media, and the pupils' exposure to the language through traditional and new media).

In the ESLC results, a positive relationship is discernible between proficiency in the tested language and the pupils' perception of their parents' knowledge of that language, and their exposure to and use of the tested language through traditional and new media.

The language-friendly school environment

Differences are found in schools' degree of language specialisation, the availability of ICT facilities, the number of guest teachers from abroad and provisions for pupils with an immigrant background. However, exchange visits for pupils, and participation in school language projects display a relatively low take-up and most aspects concerning classroom practice display relatively less variation across educational systems (such as the use of ICT for foreign language learning and teaching, the relative emphasis teachers place on particular skills or competences, emphasis on similarities between languages, and pupils' attitudes to their foreign language study, its usefulness and difficulty). Only the amount of foreign language spoken in lessons shows clear differences across educational systems.

Pupils who find learning the language useful tend to achieve higher levels of foreign language proficiency and pupils who find learning the language difficult achieve lower levels of foreign language proficiency, according to the ESLC. Also, a greater use of the foreign language in lessons by both teachers and pupils shows a positive relation with language proficiency. Overall, differences in language specialisation, hosting staff from other language communities, and provisions for immigrant pupils show no clear relationship with foreign language proficiency.

Teacher qualifications and training

Improving the quality of initial teacher education and ensuring that all practising teachers take part in continuous professional development has been identified as a key factor in securing the quality of school education in general. Overall, most language teachers are well qualified, are educated to a high level, have full certification and are specialised in teaching languages. Also, relatively little variation was found between educational systems concerning in-school teaching placements and teaching experience even though differences exist in the number of different languages teachers have taught. Generally, across educational systems only a small proportion of teachers have participated in exchange visits, despite the availability of funding for such visits in a number of educational systems. We did find considerable differences between educational systems in teacher shortages, and in the use of and received training in existing European tools, i.e. the CEFR, and the European Language Portfolio. Concerning continuous professional development, despite clear differences found in the organisation of in-service training (such as financial incentives, when teachers can participate in training and the mode of training), reported participation in and focus of in-service training display less variation across educational systems.

The different indices related to initial and continued teacher education presented in the ESLC show little relation to language proficiency. For many indices this lack of a relation can be attributed to a lack of differences within educational systems. For others however, such as the use of and received training in the CEFR, considerable policy differences have been found, and yet these differences do not account for differences in language proficiency.

Interpretation: The successful language learner

The above findings present quite a complex picture. However, a very brief summary of some of the significant findings does provide a compelling portrait of the successful language learner: a language is learned better where motivation is high, where learners perceive it to be useful, and where it is indeed used outside school, for example in communicating over the internet, for watching TV, or travelling on holiday. Also, the more teachers and students use the language in class, the better it is learned.

These conclusions are not surprising: they probably confirm what we already believed. However, it is an important achievement that the ESLC has provided empirical evidence in support of them. What the paragraph above describes is language being used for motivated, purposeful communication. It is this which favours learning: we learn in order to communicate, and we learn by communicating. Moreover, the ESLC shows that this ideal learning situation is approximated only in some countries, and effectively, only for English.

Both the language test and questionnaire results confirm that English appears as a special case. It is learned to the highest level (note in Figure 2 that it is the Flemish and German communities of Belgium which come at the top. English is their second foreign language, but they still perform more highly in it than in their official first foreign language, French). The questionnaires indicate that English stands distinct from other languages in terms of student perceptions of its usefulness, its visibility in life outside school, and its use as a medium for communication – a *lingua franca*. The successful learner of English appears to perceive and experience it in ways which are characteristically different from less successful learners of any language – including English.

Does English provide a model which other languages can follow? Clearly, it has advantages that other languages do not: above all, its higher visibility in many kinds of media. In Sweden, which as shown in Figure 1 performs very strongly in English and very poorly in Spanish (see Figure 2), English is the language of a significant proportion of television programming. This is not the case for France, which performs poorly in English. In England there is an evident credibility problem: motivation to learn a foreign language is low because it is widely perceived as unnecessary, in a world where everyone else is believed to speak English.

However, the fundamental importance of placing communication at the heart of successful language learning applies to any language. The European Commission seems inclined to acknowledge the special status of English as the language for business and for mobility (and to justify an emphasis on English in terms of pressing economic needs), while stressing the cultural importance of other languages. However, we could argue that this is not a necessary either/or choice. Effective intercultural communication goes beyond the merely utilitarian or transactional, and being able to talk to an interlocutor in their own language, even to a modest level, is an asset, in business as in social life. The ESLC provides evidence that using language in purposeful communication favours learning, and that is perhaps the most important message to take from it: languages *will* only be successfully learned as communication tools. And in the age of social networking the current generation of learners has no shortage of things to communicate about. We may assert that an appropriate language policy for Europe in the 21st century will place communication and intercultural competence at the centre. For most learners this may well start with English, but it need not and should not finish there.

The importance of the ESLC for Cambridge English Language Assessment

The wide range of achievement across countries demonstrated by the ESLC has clear implications for educational policy makers tasked with carrying forward a policy for languages. There is no European norm – every country is different, not only in terms of the parameters studied in the survey, but in other important respects: its educational traditions, the structure of its industry and business, and above all perhaps the cultural, historical and linguistic factors which contribute to the image it entertains of itself as being 'good' or 'bad' at languages – possibly powerful stereotypes with a self-fulfilling positive or negative impact. Such attitudes are clearly visible in the national press commentaries provided on the ESLC outcomes in different countries (see Stephen McKenna's article in this issue).

The ESLC serves a European policy objective of improving the effectiveness of language learning. This aligns closely with Cambridge English Language Assessment's objectives, as shown by the considerable research effort currently invested in studying the impact of its exams in particular educational contexts. *Research Notes* issue 50 presents a number of such case studies (Cambridge ESOL 2012). Many of these contexts involve collaboration with education systems at national, regional or institutional level. The work already done in the area of impact shows Cambridge English Language Assessment strongly engaged in working with countries in Europe and elsewhere on the formulation of language education policy, and assisting in its implementation.

As presented above, the ESLC outcomes can be seen to provide an intuitive recipe for success, based on the central notion of purposeful language use for communication. This is what the Cambridge English exams set out to test, tailored to different levels and learner groups. The importance of this for learning is to be emphasised.

The CEFR is used by the ESLC and Cambridge English Language Assessment as the basis for reporting and interpreting results – the historical, conceptual and empirical links between the Framework and the latter are well documented (Milanovic 2009, Taylor and Jones 2006). The CEFR is the essential point of reference which gives meaning to

measures of language proficiency. As it becomes more widely adopted, and as interpretations of its levels become more standardised across languages (a process to which the ESLC itself should contribute), so it acquires ever more meaning and utility. The implementation of the ESLC tests in five languages, in collaboration with the Association of Language Testers in Europe (ALTE) partners, shows a capacity to work with and interpret the CEFR.

At the same time, Cambridge English Language Assessment has consistently explained that reference to the CEFR must treat each context of learning on its own terms, and policy that will impact positively on language learning must be made country by country, on the basis of case studies. The great differences between contexts identified by the ESLC clearly confirm the importance of this. The purpose remains to improve learning outcomes: the notion of 'positive impact by design' (Saville 2012:5).

Learning-oriented assessment (LOA) is a current area of research within Cambridge English Language Assessment aimed at developing models for achieving positive impact by design. LOA encompasses both the familiar, formal manifestations of assessment in setting objectives and measuring achievement, as well as its informal manifestations within classroom interaction: correction and feedback, self-monitoring, reflection and evaluation. LOA seeks to define strongly complementary roles for assessment and teaching expertise. It concerns creating the *conditions* for learning, by ensuring that students have clear evidence of progress towards their goals, and are given learning tasks at an optimal level of challenge; and how to enable *mechanisms* for learning, focusing on meaningful communication activities, providing support to make new language accessible, and using every kind of individualised feedback – from tests, from the teacher, or by student self-monitoring. LOA makes no distinction between 'summative' and 'formative' because it sets out to make *all* forms of assessment useful for learning. Every assessment is an opportunity to learn, and every learning event is an opportunity to assess – that is, to evaluate, exchange feedback, record and reflect. At its centre is the notion of purposeful language use and communication.

Finally, in digesting the findings of the ESLC and considering their relevance to formulating and implementing language education policy we have an opportunity to move the focus from individual languages to the entirety of language education in a given context. We can agree with the *Guide for the Development and Implementation of Curricula for Plurilingual and Intercultural Education* (Council of Europe 2010:29) when it asserts: 'language teaching in schools must go beyond the communication competences specified on the various levels of the CEFR'.

Language education implies more than achieving some level of proficiency in a language. It comprises a range of learning skills and learning objectives that are critical to becoming competent learners not just in one language, but more importantly, given that the languages we need in later life are probably not those we learned at school, of languages generally. Though the focus of impact studies may be on English, going forward we may increasingly be treating English as just one element in a coherent comprehensive policy for language education in a given context.

Moreover, language education impacts crucially on educational outcomes generally. Hawkins (1999:138) wrote of an 'apprenticeship in languages':

> We will no longer measure effectiveness of the apprenticeship in languages by mere ability to 'survive' in a series of situations, but by how the foreign language experience contributes to learning how to learn through language, and to confidence as a (mathetic) language user.

Mathetic means: serving discovery, understanding and learning. Hawkins emphasised the importance of mother tongue competence to success in school, and of foreign languages in developing awareness of how language works.

Jones (2013) discusses the nature of the extended framework that will serve the wider conceptualisation of language education. It encompasses the concept of plurilingualism – the complex array of interconnected competences which characterises individuals who have learned or encountered more than one language. This is potentially a more productive concept for educational policy making than the current 'mother tongue plus two' target proposed by the European Commission. It provides an appropriate framework for pursuing positive impact by design.

Conclusion: The importance of criterion reference

The ESLC has given Cambridge English Language Assessment an opportunity to engage with important issues in European language education, and to contribute information for making language policy. The experience complements and extends our work on impact, which continues through bilateral collaborations in a number of countries. It helps us to conceptualise the role in language education which we should be aspiring to play, and the vision for languages in Europe which we should be aiming to promote.

Above all, perhaps, the ESLC provides an opportunity for Cambridge English Language Assessment to communicate the value of setting the right goals and using assessments which measure them to language education. It confirms the importance of focusing on communicative language ability and of setting objectives and evaluating outcomes in meaningful terms, that is, the levels of the CEFR. The presentation of countries' results in the ESLC provides a compelling example of how to focus transparently on the useful outcomes of learning.

It was not in the terms of reference for the ESLC to address the role of examinations in language education. As explained above, the questionnaires were to focus on policy issues. In any case, a country's exam system might well be a far too difficult or politically sensitive area for an international survey to address. However, to what extent the outcomes of language learning are determined for good or ill by the examinations which are set in a given educational context remains an important question.

A country for which relevant summary data is available is England. As noted, England performed very poorly in the ESLC, with almost 80% of students not achieving better than CEFR A1 in the first foreign language. And yet the cohort went on to perform well in terms of exam grades achieved in the General Certificate of Secondary Education (GCSE) language exams in 2012: 'impressive, and above the national average',

as a teacher's association called the results (Association for Language Learning 2012). Given the results one may wonder if the communicative competence which the ESLC set out to test is what English students are learning, or what the GCSE is measuring (I hope this will be the subject of a pending study, so I will leave it as an open question here). If this comparison of GCSE grades and CEFR levels were shown to be valid it would make the case strongly for re-focusing attention on the useful outcomes of learning languages, setting objectives and reporting results in meaningful terms (the CEFR) rather than in exam grades to which no meaningful interpretation can be attached.

References

Association for Language Learning (2012) GCSE results day 2012– updates, news and comment, in *Association for Language Learning*, available online: www.all-languages.org.uk/news/news_list/gcse_results_day_2012_updates_news_and_comment

Cambridge ESOL (2012) *Research Notes* 50, Cambridge: Cambridge ESOL.

Council of Europe (2010) *Guide for the Development and Implementation of Curricula for Plurilingual and Intercultural Education*, Strasbourg: Council of Europe.

European Commission (2005) *Commission Communication of 1 August 2005 – The European Indicator of Language Competence*, Brussels: European Commission.

European Commission (2012a) *First European Survey on Language Competences: Final Report*, Luxembourg: Publications Office of the European Union, available online: ec.europa.eu/languages/eslc/index.html

European Commission (2012b) *First European Survey on Language Competences: Technical Report*, Luxembourg: Publications Office of the European Union, available online: ec.europa.eu/languages/eslc/index.html

Hawkins, E W (1999) Foreign language study and language awareness, *Language Awareness* 8, 124–142.

Jones, N (2013) Defining an inclusive framework for languages, in Galaczi, E D and Weir, C J (Eds) *Exploring Language Frameworks: Proceedings from the ALTE Kraków Conference, July 2011*, Studies in Language Testing volume 36, Cambridge: UCLES/Cambridge University Press, 105–117.

Milanovic, M (2009) Cambridge ESOL and the CEFR, *Research Notes* 37, 2–5.

Saville, N (2012) Applying a model for investigating the impact of language assessment within educational contexts: The Cambridge ESOL approach, *Research Notes* 50, 4–8.

Taylor, L and Jones, N (2006) ESOL exams and the Common European Framework of Reference (CEFR), *Research Notes* 24, 2–5.

Innovation in language test development

MARTIN ROBINSON ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

The European Survey on Language Competences (ESLC) was the first survey of its kind; the first to provide empirical evidence relating to outcomes of language education across a range of European countries. A critical factor in the overall success of the ESLC was the development of high-quality, comparable language tests in five languages and successfully administering them in a large-scale, international survey.

The successful development and delivery of the language test instruments was complex, depending on the use and further development of state-of-the-art methodologies and technologies. In addition, it required the intensive collaboration among the members of the SurveyLang consortium (who were tasked with conducting the ESLC), the participating countries and the European Commission. This article briefly describes the processes adopted to develop the language tests and emphasises the aspects that were particularly innovative:

- the development of a language testing framework and construct common to the five languages
- targeted testing
- a linked test design
- a detailed test production process implemented across the five teams of language partners
- a state-of-the-art item authoring tool and item banking system
- targeted commissioning
- cross-language task adaptation
- cross-language vetting.

The first part of this article describes the language testing framework that provided the basis for the development of the language testing instruments, incorporating the aims and objectives of the ESLC. The item development process that enabled the language partners to work together in a highly collaborative and intensive way is then described, emphasising those aspects that were new and different from anything done previously.

The language testing group

The language testing group consisted of Cambridge English Language Assessment (previously known as Cambridge ESOL), Centre international d'études pédagogiques (CIEP), Goethe-Institut, Università per Stranieri di Perugia and Universidad de Salamanca. The language test production was managed by Cambridge English Language Assessment.

Development of the language testing framework

The European Commission specified the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001) as the framework against which to measure language learning outcomes for the ESLC, reflecting the widespread impact which this document has had since its initial publication. The language tests developed for the ESLC set out to reflect the CEFR's action-oriented, functional model of language use, while ensuring relevance for 15–17 year olds in a school setting. The socio-cognitive model adopted was based on the CEFR's model of language use and learning, and identified two dimensions – the social dimension of language in use, and the cognitive dimension of language as a developing set of competences, skills and knowledge. This model was used to define the testable abilities at each proficiency level to enable the resulting test construct to be implemented comparably across languages. These abilities were mapped to specific task types.

The approach to developing the language testing framework by SurveyLang is summarised as follows:

- identify the relevant aims and objectives of the ESLC, including the language skills to be tested
- for each skill, identify the test content and a set of testable subskills derived from a socio-cognitive model of language proficiency and a listing of language functions or competences found to be salient at each level from A1 to B2 in the descriptor scales of the CEFR
- identify the most appropriate task types to test these subskills
- adopt a targeted approach to testing where pupils are given a test at an appropriate level of challenge
- create a test design that presents combinations of tasks to students in such a way as to maximise the quality of interpretable response data collected while not overburdening the sampled students
- develop specifications, item writer guidelines and a collaborative test development process that are shared across languages in order to produce language tests that are comparable.

Aims and objectives of the ESLC

The aim of the survey was to deliver an indicator of language competences to provide information on the general level of foreign language knowledge of the pupils in the Member States in order to help policy makers, teachers and practitioners to take decisions on how to improve the foreign language teaching methods and thus the performance of pupils.

The aim of the SurveyLang language testing group was to develop language tests, the results of which were comparable across the five languages and all participating countries. This broad aim could be broken down into a number of key objectives which impacted on the design of the language testing instruments:

- for each country, the ESLC should cover tests in the first and second most commonly taught official European languages in the European Union from English, French, German, Italian and Spanish
- test performance should be interpreted with reference to the CEFR scale
- the tests should assess performance at Levels A1–B2 of the CEFR
- performance should be reported at the level of the cohort, not the individual
- the ESLC should assess competence in the three language skills which may be assessed most readily, i.e. listening comprehension, reading comprehension and writing
- results must be comparable across the five languages and all participating countries
- tests must be available in both paper-based (PB) and computer-based (CB) formats
- testing time at the individual level must be kept to a minimum; at the same time, the reliability and validity of the data must be maximised at the cohort level
- instruments for testing in the three skills should be developed taking into account the previous experience and knowledge in the field at international, European Union and national level.

Previous international surveys had translated tests across languages but it was a key aim of this survey to create parallel but not identical tests across the five languages, thereby making the issue of cross-language comparability a crucial one.

Test content and subskills to be tested

Test content was approached using the domains of language use proposed by the CEFR (Council of Europe 2001:43–100). As the CEFR stresses, these categories are illustrative and suggestive, rather than exhaustive. However, the listed elements provided a useful starting point for selecting appropriate content.

The CEFR identifies four basic domains of language use:

- personal
- public
- educational
- professional.

The CEFR illustrates each domain in terms of situations (e.g. the locations in which they occur), communication themes (e.g. daily life) and topic-specific notions (e.g. family celebrations and events, relationships, etc.). Consideration of which domains of language use are most relevant to target language learners at different proficiency levels informed a decision as to the proportion of tasks relating to each of the domains mentioned above across the four levels of the ESLC. This distribution is illustrated in Table 1.

Table 1: Domain distribution across Levels A1–B2

	A1	A2	B1	B2
Personal	60%	50%	40%	25%
Public	30%	40%	40%	50%
Educational	10%	10%	20%	20%
Professional	0%	0%	0%	5%

Language functions (e.g. imparting and seeking information) are discussed in the CEFR as an aspect of pragmatic competence, in order to provide a general rather than setting-specific taxonomy of language in social use.

Together, these communication themes, notions and functions provided the basis for categorising and selecting texts for use in the ESLC. The choice of test content also took into account the characteristics of the target language users, i.e. the 15–17 year old students participating in this survey. To ensure adequate coverage across the ESLC, domains and topics were assigned to tasks at the commissioning stage.

Task types

The socio-cognitive validation framework proposed by Weir (2005), an approach consistent with other more recent discussions of theories of test design, was adopted as the means to identify the subskills to be tested. This complements the CEFR's treatment of the cognitive dimension and provides useful practical models of language skills as cognitive processes. Testable abilities at each proficiency level were defined using the CEFR and socio-cognitive framework. In order that the resulting test construct should be implemented comparably across languages, these abilities were mapped to specific task types, drawing chiefly on task types used successfully by the consortium's language partners in their exams. A rigorous design was proposed which could be replicated across languages, thus maximising coherence and consistency in the implementation of the construct.

For reading and listening it was preferred to use selected response types, for ease and consistency of marking:

- multiple-choice (graphic options, text options, true/false)
- multiple-choice gap-fill (gapped texts, e.g. to test lexico-grammatical relations)
- matching texts to graphics (e.g. paraphrases to notices)
- matching texts to texts (e.g. descriptions of people to a set of leisure activities/holidays/films/books that would suit each of them)
- matching text elements to gaps in a larger text (e.g. extracted sentences) to test discourse relations, understanding at text level.

For writing, a range of open, extended response task types was used in keeping with the CEFR's action-oriented, communicative, functional model of language use, e.g. writing an email, postcard or letter, or writing a referential or conative text (intended to inform, persuade or convince).

Targeted testing

A targeted testing approach was adopted to ensure that, as far as possible, students were presented with test items of an appropriate level of difficulty, i.e. students at lower levels of ability were not presented with items that were too difficult and students at higher levels were not presented with items that were too easy. This would minimise the demotivating aspect of having to deal with inappropriate items while simultaneously maximising the reliability of the data.

The targeted approach necessitated the administration of a routing test for each language during the sampling process in advance of the test date. Each routing test was 15 minutes long, and for simplicity consisted of 20 reading-focused items, ordered to be progressive in difficulty. On the basis of the results of the routing test, students were placed into one of three broad level groups and received a test of low, medium or higher difficulty.

Test design

Unlike in standard language testing where the focus is on the individual, in complex surveys such as the ESLC each sampled student need only see some of the total test material. The total amount of test material was determined by the need to achieve adequate coverage of the construct, i.e. to test all aspects of a skill considered important at a given level. In order to avoid fatigue or boredom among students it was necessary to utilise an incomplete but linked design where each student would receive only a proportion of the total test material.

A design constraint was adopted that the total language test time for a student should not exceed 60 minutes. A test for one skill would comprise 30 minutes of material. A student would only be tested in two of the three skills. Individual students would therefore receive reading and listening; reading and writing; or listening and writing. Students would be assigned randomly to one of these three groups.

The targeted testing described above would ensure that students were placed into one of three broad level groups but the short routing test could not be accurate enough to assign an exact CEFR level. Each test would therefore need to cover two CEFR levels. A complex design of test booklets with overlapping content would also allow for the same task to be placed in different positions in different tests to negate any potential task order effect. To facilitate the implementation of a linked test design, all tasks were constructed with time loads of 7.5, 15 or 30 minutes.

Figure 1 indicates how students were placed into one of three broad level groups according to the results of the routing test. The test at each level spanned two levels of the CEFR. Figure 2 illustrates the linked test design in more detail for one of the three skills, reading.

Figure 1: Targeted test design

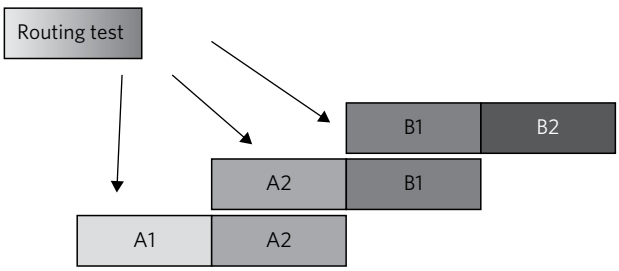


Figure 2: Linked test design for reading

Level 1												
Tasks	**English**	**French**	**German**	**Italian**	**Spanish**	**Time**	**Booklet 1**	**Booklet 2**	**Booklet 3**	**Booklet 4**	**Booklet 5**	**Booklet 6**
A1-R1	ER112	FR112	GR111	IR113	SR112	**7.5**	**1**	**2**		**2**	**1**	
A1-R2	ER211	FR211	GR213	IR211	SR211	**7.5**	**2**		**1**	**1**		**2**
A1-R3	ER312	FR311	GR312	IR313	SR312	**7.5**		**1**	**2**		**2**	**1**
A2-R2	ER223	FR223	GR221	IR223	SR223	**7.5**	**3**		**4**		**3**	
A2-R3	ER321	FR322	GR321	IR323	SR322	**7.5**	**4**	**3**				**4**
A2-R4	ER423	FR423	GR421	IR421	SR423	**7.5**		**4**	**3**	**4**		
A2-R5	ER523	FR523	GR522	IR521	SR523	**7.5**				**3**	**4**	**3**
						Total time	**30**	**30**	**30**	**30**	**30**	**30**
Level 2												
Tasks	**English**	**French**	**German**	**Italian**	**Spanish**	**Time**	**Booklet 7**	**Booklet 8**	**Booklet 9**	**Booklet 10**	**Booklet 11**	**Booklet 12**
A2-R2	ER223	FR223	GR221	IR223	SR223	**7.5**	**2**		**1**		**2**	
A2-R3	ER321	FR322	GR321	IR323	SR322	**7.5**	**1**	**2**				**1**
A2-R4	ER423	FR423	GR421	IR421	SR423	**7.5**		**1**	**2**	**1**		
A2-R5	ER523	FR523	GR522	IR521	SR523	**7.5**				**2**	**1**	**2**
B1-R5	ER532	FR531	GR533	IR531	SR531	**7.5**	**3**	**4**		**4**	**3**	
B1-R6	ER631	FR631	GR633	IR632	SR631	**7.5**	**4**	**3**	**3**			**4**
B1-R7	ER731	FR733	GR731	IR733	SR733	**7.5**			**4**	**3**	**4**	**3**
						Total time	**30**	**30**	**30**	**30**	**30**	**30**
Level 3												
Tasks	**English**	**French**	**German**	**Italian**	**Spanish**	**Time**	**Booklet 13**	**Booklet 14**	**Booklet 15**	**Booklet 16**	**Booklet 17**	**Booklet 18**
B1-R5	ER532	FR531	GR533	IR531	SR531	**7.5**	**1**	**2**				
B1-R6	ER631	FR631	GR633	IR632	SR631	**7.5**		**1**	**1**			
B1-R7	ER731	FR733	GR731	IR733	SR733	**7.5**	**2**		**2**			
B2-R6	ER642	FR642	GR642	IR642	SR641	**15**			**3**	**1**		**2**
B2-R7	ER741	FR743	GR741	IR743	SR741	**15**	**3**			**2**	**1**	
B2-R8	ER841	FR843	GR842	IR842	SR841	**15**		**3**			**2**	**1**
						Total time	**30**	**30**	**30**	**30**	**30**	**30**

In Figure 2 the first column indicates the CEFR level (A1–B2) and the task type (there are eight reading task types, R1-R8). Columns 2 to 6 indicate the language and the task ID of each task while the seventh column indicates the time load for the task in minutes. Columns 8 to 13 indicate how each task was placed into different tests or booklets. For example, booklet 1 appears in column 8. This is a level 1 booklet containing four tasks, each of 7.5 minutes. The numbers 1–4 represent the order of the tasks so that task 1 (A1-R1) in booklet 1 becomes task 2 in booklet 2. Each task appears in a number of booklets. Booklet 1 has two A1 level tasks and two A2 level tasks. The fourth task, A2-R2, not only appears in other level 1 (A1/A2) tests but also appears in level 2 (A2/B1) tests, e.g. booklet 7.

It can be seen from this design that each individual test or booklet only consisted of three or four tasks and only lasted 30 minutes. However, the complete design ensured that the whole construct, i.e. all task types, was tested at the cohort level.

The design was implemented in the same way in each of the five languages, as consistency of approach would maximise the comparability of outcomes.

Test development

As stated above, the aim of the language testing group was to develop language tests, the results of which were comparable across the five languages and all participating countries. To achieve such high levels of comparability and quality required a high degree of collaboration from the language testing group and the adoption of a shared:

- test development cycle (from initial pilot through to Main Study)
- test construct, specifications, item writer guidelines
- test production process
- item authoring tool and item banking system
- quality control process.

The approach adopted by SurveyLang in designing the language test instruments is summarised as follows:

(i) Define a language testing framework that incorporates the aims and objectives of the ESLC (described in the previous section *Test content and subskills to be tested*).

(ii) Out of this framework, develop initial specifications, a set of draft task types and a draft test development process.

(iii) Pilot the initial specifications and draft task types.

(iv) Gather feedback from all relevant stakeholders including the European Commission, the participating countries, teachers and students. Review this feedback together with the analysis of the pilot results.

(v) Further develop the initial specifications into final item writer guidelines and agree on a collaborative test development process to be shared across the five languages.

(vi) Undertake a rigorous item development programme.

Test development cycle

There were five main stages in the development of the language testing instruments, which can be summarised as follows:

- development of the language testing framework (2008)
- the Pilot Study (2008)
- Pretesting (2009)
- the Field Trial (2010)
- the Main Study (2011).

To ensure that the items used in the Main Study were fit for purpose and of the required level of quality, the language testing team produced and trialled a large number of items over the course of the development programme. Over 100 tasks were piloted in 2008 in order to finalise the test specifications and obtain the agreement of participating countries on the most appropriate task types to be used in the ESLC. The team then produced over 500 tasks (2,200+ items) which were then exhaustively trialled through the Pretesting and Field Trial stages before the best-performing items were selected. For the Main Study, 143 tasks (635 items) were used across the five languages.

Each of these development stages contributed to the specification of the tests, in terms of content and task types, to the construction of a large body of test tasks, and to their progressive refinement through a series of empirical trials and the collection of qualitative feedback.

An important conditioning factor was the collaborative working methodology itself, developed by the language partners in order to maximise the quality and the comparability of the final tests.

Test construct, specifications and item writer guidelines

Following the Pilot Study, the test specifications were reviewed and finalised. Common test specifications across the five languages ensured that tasks across languages were almost identical in terms of number of items, number of options, text length, etc.

Detailed item writer guidelines were developed for each of the three skills. These guidelines specified the requirements of each task type at each level in terms of overall testing aim, testing focus, level of distraction in the options, input text length, etc. They also provided explicit guidance on the selection and manipulation of text types and topics, and the production of artwork and recordings. Quality criteria relevant to each task type were listed and these criteria provided the basis for the acceptance, rejection and editing of tasks as they proceeded through the item production process.

Item authoring tool and item banking system

Close collaboration between the partners in the development of the tests, and consistent implementation and presentation of test tasks, were supported by the item authoring, banking and test assembly functionality of the testing tool specifically developed for the ESLC by the responsible partner, Gallup Europe, with input from the language partners.

The development provided an integrated, state-of-the-art, functionality-rich software system for the design, management and delivery of the language tests. The platform was fine-tuned to the specific set of requirements of the ESLC project and was designed to support the delivery of PB and CB tests. The software platform also supported all major stages of the survey process.

The test items of the ESLC were developed by an expert team of 40+ item writers distributed across Europe, who worked according to specifications and guidance provided by the central project team. Items were moved through various stages of a predefined life-cycle including authoring, editing, vetting, adding of graphics and audio, pilot-testing, Field Trial etc. Each stage involved different tasks, roles and responsibilities.

The test-item authoring tool was designed to support this distributed and fragmented development model. It was also designed to allow non-technical personnel to create tasks in an intuitive way by means of predefined templates for the various task types that were used in the survey. At any stage in the development, a task could be previewed and tested to allow the author to see how it would look and behave when rendered in a test. The authoring tool also supported the capture and input of all the metadata elements associated with a task, including descriptions, classifications, versioning metadata, test statistics, etc.

The tool was implemented on candidate computers by means of technologies including Adobe Flex and Adobe Air. This provided a user-friendly and aesthetically pleasing environment for the various groups involved in the development of the tasks.

As an integrated part of the life-cycle system, the functionality to create new versions of and adapt tasks was implemented. When a new version of a task was created, any changes to it would only affect the latest version. Adaptation was, on the other hand, a procedure that allowed a task developed in one test language to be adapted to another language.

One of the most innovative features of the Item Bank was its ability to manage the audio tracks of the listening tasks. Creating high-quality audio is normally a time-consuming and expensive operation. Traditionally the full length track of a task is created in one go and stored as an audio file. If a change is made to this task at a later stage, the audio file is no longer usable and a completely new recording is thus required. Furthermore, a test-length recording that records each task twice and also records the silences creates an unnecessarily large audio file. To avoid this, an audio segmentation model was developed whereby

the audio files could be recorded as the shortest possible audio fragments. The various fragments were stored along with the other resources of the task and were only assembled into full-length audio tracks at the point of test assembly.

By using a shared, integrated, online software system for the production of the language tests, the language testing group, although dispersed across Europe for the duration of the project, could ensure that each language team and all of its members were following the same procedures at the same time.

Test production process

The successful delivery of the language test instruments depended heavily on a shared, collaborative test production process that was detailed and comprehensive enough to produce items of the required high levels of quality. The production of items that were comparable across the five languages additionally required some innovative methodologies.

The steps in the test development process are shown in detail in Figure 3. It can be seen that this is a very detailed, complex process. While most stages (item writing, editing, pretesting, etc.) will be familiar to language testers, what makes this process unique are the additional stages of targeted commissioning, cross-language vetting and cross-language adaptation.

Targeted commissioning

Before item writing began, the number of items required for the Main Study was calculated. As the pretesting and Field Trial stages were intended to enable selection of the best performing items for the Main Study, a much greater number of items than required for the Main Study were therefore commissioned. In total, over 500 tasks (2,200+ items) were commissioned across the five languages. Given the large number of item writers commissioned, it was imperative to plan for adequate coverage of construct, domains and topics for all tasks at each level across the five languages. Each item writer therefore received a detailed commissioning brief specifying the task types, levels and topics to ensure adequate and consistent coverage of the CEFR domains.

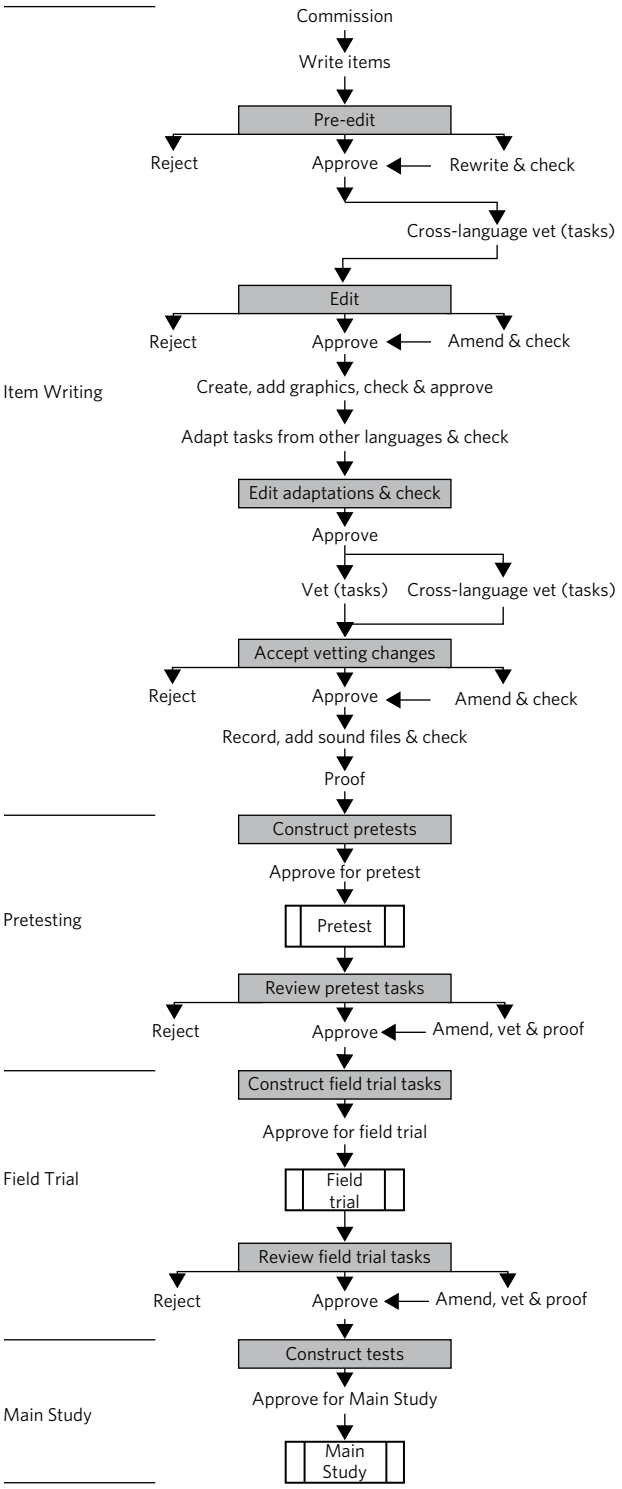
The work of creating these tasks was divided among the language partners according to the strengths of each item writing team. Over 40 specialist item writers were commissioned across the five languages. For some languages, item writers specialised in certain skills, levels or task types. Item writers were organised into teams and managed by team leaders and specialist language testing product managers.

Cross-language task adaptation

The common approach to item development described in Figure 3 was considered essential to ensure comparability over the five languages in the way they related performance to the CEFR.

Task adaptation worked as follows. In the Pilot Study, a proportion of the tasks were adapted across languages. Each

Figure 3: Test development process



language partner was asked to adapt some tasks from two of the other four languages. There were several purposes for adapting tasks:

- it was seen as a valuable context for developing collaborative working methods between the language partners: studying each other's tasks in detail stimulated much critical reflection and interaction
- it might be a possible way of enhancing consistency and comparability across languages
- it might offer a straightforward, if not a quicker, way of generating new tasks.

The Pilot Study review confirmed the value of adapting tasks across languages. It appeared that most task types used in the Pilot Study could be successfully adapted from one language into another if the aim was to adapt but not translate. However, the process needed skilled item writers who were not only competent in two or more of the languages, but also had a detailed and comprehensive understanding of the CEFR and its language descriptors. Item writers needed to be aware of lexico-grammatical differences between the languages and how these differences might affect the perceived difficulty of the items. The only task type that appeared difficult to adapt was the multiple-choice cloze task where the testing focus was largely lexico-grammatical. It was discovered that adaptation was most practical at the lower levels and although possible with some higher-level task types, the longer texts involved meant the extra effort required sometimes outweighed the benefits. For the skill of writing, no significant difficulties were encountered with adapting any of the task types.

Thus, it was deemed practical and desirable to adapt all the writing tasks at all levels and all reading and listening tasks at A1 and A2. This was taken into account at the commissioning stage where each partner only needed to write a proportion of the required writing tasks and reading and listening tasks at A1 and A2.

Cross-language vetting

As well as cross-language task adaptation, cross-language vetting was another innovative addition to the ESLC test production process that was considered to have significantly beneficial effects.

Cross-language vetting worked as follows:

- tasks from each language were vetted by at least two other language partners, again according to the language strengths of the members of each partner's team
- multilingual, experienced item writers vetted tasks from other languages to ensure that tasks, items and options would operate correctly
- a vetting form was created to ensure that vetting comments could be recorded consistently and electronically
- vetting comments were then passed back to the original language partner who could then compare comments from both their own vetters and the vetters from other language partners.

This methodology was trialled during the Pilot Study and a review conducted at the end of that stage confirmed the value of cross-language vetting as an additional stage to the standard test production process. It not only provided a valuable additional quality control, it also enabled the sharing of knowledge and experience among the language partners.

Conclusions

The production of the language testing instruments took four years and many valuable lessons were learned. At the end of each stage in the developmental process, the language partners reviewed and refined not only the test materials themselves but also the methodologies used to create them. It is clear from the success of the ESLC and the quality of the test materials produced that some of the innovative techniques introduced played a valuable part in the production process and should be seen as approaches to be taken forward and further developed. In particular, techniques such as cross-language task adaptation and cross-language vetting may have the potential for wider application in multilingual language test production contexts, and possibly in any potential second round of the ESLC. It should be emphasised, though, that adaptation is not translation. Its potential will only be achieved through its use and development by trained and experienced item writers who are multilingual and have a comprehensive understanding of relevant languages, language itself and the CEFR.

Finally, it should be noted that the successful development of the language tests, being methodologically complex and extremely challenging, was only made possible by the intensive collaboration, the willingness to adapt and the openness to innovative ideas demonstrated by the members of the language testing group.

References

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

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

The European Survey on Language Competences – the Polish experience

MAGDALENA SZPOTOWICZ EDUCATIONAL RESEARCH INSTITUTE, POLAND

Introduction

Poland was one of 14 participating countries in the 2011 European Survey on Language Competences (ESLC) (European Commission 2012). The survey results are a rich source of information about language learning and teaching and also provide a solid base for further investigation and

study. Primarily the results provided valuable information about the Polish secondary-school learners' language competence measured on the Common European Framework of Reference for Languages (CEFR) scale (Council of Europe 2001), which is now a point of reference in the new core curriculum (Ministerstwo Edukacji Narodowej 2008). The contextual data collected in the survey created the opportunity for detailed insight into the teaching and learning of foreign languages. This highlighted strengths and potential areas for improvement, and findings are available in the national ESLC report (Instytut Badań Edukacyjnych in press).

This first ever external measurement of language competences on a representative sample of learners was a unique opportunity to compare learning outcomes from the Polish educational system with those from other European countries. It was further used to directly compare the same students' results in the ESLC with national language exams. This showed a similar distribution of the results in the ESLC and national foreign language (FL) exam results and a relatively high correlation between them (Kulon, Gajewska-Dyszkiewicz and Paczuska 2012). In addition to the main ESLC suite, in Poland alone, an oral test was co-administered to a sub-sample (N = 499). This test aimed to measure spoken competence which was otherwise not tested in the ESLC. The oral results demonstrated the low attainment of many Polish learners and there was a positive association between the oral test results per CEFR level and the results of the written component of the ESLC.

It needs to be stated that the survey contributed to the promulgation of high standards in research and testing methodology in the teams who administered, monitored and analysed the data, and the wider audience and readership of subsequent publications. Reflections on improvements which could be adopted in the second round of the study will be shared in the final section of the text.

Prior to discussion of the test results and selected factors in foreign language learning, it is relevant to present the Polish educational context, especially in the light of recent curricular reforms.

Language learning environment – before and after the reform

The survey captured the last year of students who followed the 'old' curriculum (Ministerstwo Edukacji Narodowej i Sportu 2002). The ESLC participants' FL learning environment therefore differed from that of learners who are currently learning foreign languages in Polish schools. The reform of 2009 considerably changed the system of foreign language instruction by introducing three major modifications: compulsory foreign language instruction starts with the onset of schooling in grade one (ages 6–7); two foreign languages are now compulsory in lower-secondary school (ages 13–15), and continuation of the same FL from primary to lower-secondary school is required. The ESLC cohort, being the last year before the reformed curriculum, started compulsory foreign language instruction in grade four (age 10). Now all children must start in grade one. The language being studied in primary school is continued in lower-secondary to ensure a consistent period of nine years of learning a FL before taking a national exam at the end of lower-secondary school. For the ESLC cohort this exam was less demanding than it is now. It assumed a shorter (6-year) period of study and covered one basic level, which might have been demotivating for higher-achieving students. Now, exams are offered at the basic and more advanced levels, the latter being obligatory for those who continue learning the language after primary school. It needs to be noted that before the reform there was no such requirement and the ESLC-participating students might have begun their FL learning at their lower-secondary school and then had only three years of instruction at the moment of ESLC testing.

The old curriculum, which the ESLC cohort followed, did not contain any references to the CEFR; however, the development of communicative language competence was already being emphasised. The new curriculum introduced explicit references to learner achievements at the end of every stage of education. After lower-secondary school, when the ESLC tests were administered, the level has been defined as 'approaching A2' for students beginning a FL at this stage and at A2+ for students continuing their FL learning from primary school (Ministerstwo Edukacji Narodowej 2009). It means that both the students and the teacher preparing for the national exam after lower-secondary school have clearer goals and are more aware of the demands formulated in CEFR language. This is important as it will allow the national test results to be more easily compared to other measurements like the ESLC results, which are based solely on CEFR level descriptors. The second round of the ESLC will provide a more accurate comparison to the reformed exam – based on the new CEFR-referenced curriculum.

Carrying out the first international test of language competences in the last year of the old curriculum creates an opportunity to evaluate the effectiveness of the reformed system of foreign language education, provided that the second round of the ESLC captures the learners who have followed their full nine years in the reformed system from the onset of their school education. It is expected that the results will then be considerably better than those presented below.

Polish students' language test results

In Poland (as in each ESLC country) a representative sample of students participated in the survey and sat a test in one of the two most-studied foreign languages. In Poland, these languages were English and German. A cohort of about 1,700 students were tested in English and about 1,500 in German. The tests were administered in the third grade of lower-secondary school (gimnazjum) (ca. 15 year olds) and covered approximately 3,200 students, 400 teachers and 120 school principals from 146 schools.

As presented in Table 1, more than half of the students participating in the ESLC who learned English as a target language attained the CEFR Level A1 or below. Between 22.6% and 28% of the learners achieved higher results, i.e. Level B1 or B2 per skill. The percentage obtaining a result of A2, the level now required by the curriculum at this stage of education, ranged from 11.1% for reading and 15.2% for listening to 23.2% for writing.

Table 1: Percentage of Polish students achieving each CEFR level, by skill in English (the first most widely taught target language) (Instytut Badań Edukacyjnych in press)

Language skills	Pre-A1	A1	A2	B1	B2
Reading	27.1	38.1	11.1	10.3	13.4
Listening	27.4	29.4	15.2	14.6	13.4
Writing	18.7	35.5	23.2	18.8	3.8

The proportion of learners who achieved results at the level of A1 or below was even higher for those who learned German as a target language. The joint percentage of A1 and pre-A1 results, as presented in Table 2, amounted to 83.2%–87% depending on the skill. On the other hand, the percentage of students who achieved high scores (B1 and B2) was smaller than for English and covered jointly only 5.3% to almost 7% of all test results for German. At A2, the level required by the new curriculum, between 7% and 10% of students obtained these scores.

Table 2: Percentage of Polish students achieving each CEFR level, by skill in German (the second most widely taught target language) (Instytut Badań Edukacyjnych in press)

Language skills	Pre-A1	A1	A2	B1	B2
Reading	41	46	7	3.6	2.4
Listening	44.7	41.1	8.9	3.8	1.5
Writing	44.8	38.4	9.9	4.7	2.2

These results should be interpreted not only in the context of the old curriculum but also in the wider social context of foreign language learning and exposure to target languages in Poland. This context will be highlighted in the presentation of factors identified to have influenced these language results.

Factors influencing language proficiency levels

Contextual data concerning student language learning and teaching backgrounds, which was collected in the ESLC via student, teacher and school principal questionnaires, allowed for further analyses of Polish students' results to identify factors which may influence their proficiency levels. In the analyses carried out by the Polish ESLC team several factors have been identified as significant.

These findings have been divided into three groups: home-related, school-related, and wider environmental factors. Each group will be briefly discussed below.

Home-related factors

Apart from the influence of parent socio-economic status on student achievements, also observed in other studies, such as the Programme for International Student Development (PISA) (Organisation for Economic Co-operation and Development 2010), the survey revealed another important factor. Polish parents' proficiency in the target language, as perceived by students, appeared to be one of the lowest among the 14 ESLC countries. Over 80% of children estimated that their parents knew the target language only a little or not at all. This factor also appeared to influence students' level of language competence.

School-related factors

The analysis of the survey data on classroom language revealed that time spent using the target language during lessons as declared by both teachers and students was among the lowest of all countries participating in the survey. Although typical foreign language classes are groups of 11–15 or 16–20 students, the learners worked individually more often than in groups. It is important to consider this finding in light of other findings. According to the ESLC results, the intensity of communication in the target language during lessons showed a positive relationship with the students' language competence levels. It seems, therefore, desirable to encourage more target language use during language lessons.

Factors related to wider exposure and contact with target language

For most Polish students, exposure to the target language is limited to school and the media. The percentage of students who declared regular home use of the target language was minimal. Over one third of students declared they had contact with the target language through friends who communicated with them orally or in correspondence. On the other hand, media in Poland offers little exposure to original language versions of films or TV programmes. Public broadcasting delivers content with voice-overs for adults and dubbing for children.

Conclusions

Reforms implemented in Poland, as outlined above, should reap rewards in the near future and be observed in the second round of the ESLC study. However, action needs to be taken in areas not influenced by the systemic school reform.

Comparison of Polish results with those of other European contexts showed that countries where students' contact with the target language is more extensive performed better in the tests. Results of this survey should encourage:

- in schools: increased language use in meaningful situations in the classroom and with peers from schools abroad, e.g. international school projects
- in the media: broadcasting TV programmes and films in original language versions with subtitles, especially for young learners and teenagers.

This plea should be addressed to national policy makers, school principals as well as teachers and parents to promote foreign language learning in informal, everyday situations and contexts.

Implications for the second round of the survey

Involvement in the ESLC project extended over a period of three years. The experience gathered from following the rigorous procedures involved in translating instruments, administering field trials, executing the main study and finally analysing the national data has been extremely valuable and reflections on this experience have been used to base the shape and organisation of the second round of the survey. It should be recommended that:

1. All four language skills should be tested. Testing oral skills creates a number of additional challenges, so steps should

be taken as soon as possible to begin work on procedures and instruments.

2. All test takers should sit all three skills in the written test – this would simplify administrative procedures and make it easier to use the results for national linking and referencing to the CEFR, and to inform schools about the results.
3. Contextual data should be collected and coded in a way which allows individual student and teacher data to be linked – this would provide better insight into classroom practice.
4. A mixed-method approach to research should be considered, so that qualitative data could support quantitative findings (e.g. observations on sub-samples).

Finally, it should be stated that the second round of the ESLC will be an important opportunity to monitor language policy implementation in Poland as, similarly to the PISA study, it will provide an opportunity to observe change during a critical period.

References

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

European Commission (2012) *First European Survey on Language Competences: Final Report*, Luxembourg: Publications Office of the European Union, available online: ec.europa.eu/languages/eslc/index.html

Instytut Badań Edukacyjnych (in press) *Europejskie badanie kompetencji językowych. Raport krajowy: 2011* [*The European Survey on Language Competences. The national report: 2011*], Warszawa: Instytut Badań Edukacyjnych.

Kulon, F, Gajewska-Dyszkiewicz, A and Paczuska, K (2012) Porównanie wyników Europejskiego Badania Kompetencji Językowych i egzaminu gimnazjalnego z języków obcych [Comparing results of the European Survey on Language Competences and the national exam in foreign languages], in Niemierko, B and Szmigel, M K (Eds) *Regionalne i lokalne diagnozy edukacyjne* [*Regional and Local Diagnoses in Education*], Kraków: PTDE, 433–451.

Ministerstwo Edukacji Narodowej (2008) *Rozporządzenie Ministra Edukacji Narodowej z dnia 23 grudnia 2008 r. w sprawie podstawy programowej wychowania przedszkolnego oraz kształcenia ogólnego w poszczególnych typach szkół. Dz.U. nr 4 z dn. 15 stycznia 2009* [*The Regulation of 23 December 2008 by the Minister of National Education on the core curricula for preschool education and general education in individual types of school*], Warszawa: Kancelaria Prezesa Rady Ministrów.

Ministerstwo Edukacji Narodowej (2009) *Podstawa programowa z komentarzami (tom 3. Języki obce w szkole podstawowej, gimnazjum i liceum)* [*The core curriculum with comments volume 3, Foreign languages in primary, lower-secondary and upper-secondary schools*], Warszawa: Ministerstwo Edukacji Narodowej, 61–72.

Ministerstwo Edukacji Narodowej i Sportu (2002) *Rozporządzenie Ministra Edukacji Narodowej i Sportu z dnia 26 lutego 2002 r. w sprawie podstawy programowej wychowania przedszkolnego oraz kształcenia ogólnego w poszczególnych typach szkół (Dz. U. z 9 maja 2002 r. Nr 51, poz. 458)* [*The Regulation of 26 February 2002 by the Minister of National Education and Sport containing the core curricula for preschool education and general education in individual types of schools*], Warszawa: Kancelaria Prezesa Rady Ministrów.

Organisation for Economic Co-operation and Development (2010) *PISA 2009 Results: Executive Summary*, Paris: OECD Publishing, available online: www.oecd.org/pisa/pisaproducts/46619703.pdf

The European Survey on Language Competences in Croatia: Results and implications

JASMINKA BULJAN CULEJ NATIONAL CENTRE FOR THE EXTERNAL EVALUATION OF EDUCATION, CROATIA

Introduction

This article features an analysis of the results of the European Survey on Language Competences (ESLC) for Croatian students and the implications of the survey findings for policies on foreign language instruction in Croatia. The results of the survey, along with results of other research projects, show that several contextual factors have positive and significant effects on students' test results. Some of these factors are: early learning of foreign languages, learning more foreign languages, parents' knowledge of the target language, exposure to the target language through media, use of the target language during lessons, and the perception of usefulness of the target language (see Jones's article in this issue for more detail of the ESLC results). The aim of this article is to explain the relationship between the number of years students learn English and their performance on the ESLC test, and to comment on the results of Croatian students regarding other factors influencing language proficiency.

Literature review

Early foreign language learning in Croatia

A particularly interesting and important area of foreign language education is the early learning of foreign languages, starting from kindergarten, preschool or the first grade of primary school. In order to improve foreign language skills of primary school students, since 2003, every primary school student in Croatia starts learning English or German as a compulsory subject from the first grade (ages 6 or 7).

The decision regarding the implementation of early foreign language teaching in Croatia was partially based on

results from research projects conducted in Croatia. A pilot project of early English instruction was introduced in several primary schools in Zagreb in 1973. According to the project coordinator, the late Professor Vilke (1993:10), 'the ultimate aim of introducing early learning of a foreign language was the production of competent bilingual speakers throughout the country'. Vilke (1993:11) argues that the project was devised on the basis of neurophysiological and psycholinguistic evidence and empirical data obtained from language learning programmes in a variety of countries and cultures. Furthermore, Bartolović (1993:43) who investigated early foreign language learners' cognitive development, found that 'learning a foreign language did not have a negative effect on learning other subjects, because the results of those pupils were equal to or higher than the results of pupils in parallel forms'.

The project was reintroduced in 1991 and this time foreign language instruction involved three more languages: German, French and Italian. It proved successful in that it empirically corroborated the assumption that 'children of 6+ can learn foreign languages in a school environment provided teaching is shaped according to the psychomotoric and intellectual requirements of this complex age' (Vilke 1995:1). Three years later, Mihaljevic Djigunovic (1995) conducted a follow-up study concerning the attitudes towards foreign language learning on the same sample of students. The study found that the positive attitude towards game activities in class 'extended to most classroom events and that children developed an increased ability to evaluate their own proficiency and determine the benefits of learning foreign languages' (Mihaljevic Djigunovic 1995:31). Therefore, students' favourable views on foreign languages served as an additional reason to introduce early foreign language learning in the first grade of primary school.

Other factors influencing foreign language instruction

Findings from language immersion programmes in other countries support the decision to introduce foreign languages at an early age. For instance, a recent study in Greece found that the immersion of kindergarten children in an early English project 'had a positive effect on the kindergarten children's oral skills' (Griva and Sivropoulou 2009:79). An early start has also been shown to be beneficial in the long run. Domínguez and Pessoa (2005:474) found that 'early immersion students typically retain an advantage on communicative tests of listening comprehension and speaking when compared with late immersion students'.

Longer exposure supports language development, but other factors, such as lesson time and motivation, need to be carefully managed for the desired outcomes to be realised. For instance, Marinova-Todd, Marshall and Snow (2000:28) argue that: 'Children who study a foreign language for only a year or two in elementary school show no long-term effects; they need several years of continued instruction to achieve even modest proficiency'. The lack of motivation and continuity may represent a serious problem in foreign language education, even with young learners. Enever (2009:37) points out that the ELLiE (Early Language Learning in Europe) project has collected evidence 'revealing the difficulties of maintaining continuity of quality in some contexts'. In this regard, Enever (2009:38) mentions that the implementation of foreign language teaching policies should be viewed not only in the light of linguistic evidence (such as the Common European Framework of Reference (CEFR)) but also regarding social and economic issues.

Mihaljevic Djigunovic (1995:31) argues that positive attitudes and motivation may be influenced by an increased support from parents and friends. According to the ESLC results, 'more parental target language knowledge goes with higher scores on the language tests' (European Commission 2012:57). Since there already is some evidence that motivation and attitudes affect foreign language acquisition (Dörnyei 1998:117), it would be worthwhile to conduct a large-scale survey on students' self-perception and motivation. It is more difficult to directly influence contextual factors such as exposure and parental support, but creating a favourable climate for learning foreign languages would certainly prove beneficial. In regard to this, Clark (2002:184) argues that: 'When children do not have many opportunities to use language and have not been provided with a rich experiential base, they may not learn to function well in their second language, and at the same time, they may not continue to develop their first language'. However, parents' knowledge of the target language and language use at home both have 'a positive effect on the respondents' Listening and Reading scores' (European Commission 2012:58). Similarly, teachers' and students' target language use during lessons has positive effects on test results (European Commission 2012:64). Apparently, Croatia is one of the countries where teachers report frequent use of the target language (European Commission 2012:64). However, this is not in accordance with the results of the ELLiE study, which indicate that 'the lowest teacher talking time in the target language was found in Croatia' (Szpotowicz, Mihaljevic Djigunovic and Enever 2009:149). It should be noted that the ELLiE research project included first grade teachers, whose attitude to the use of the target language during lessons might be affected by the first graders' lower level of proficiency.

Results and discussion

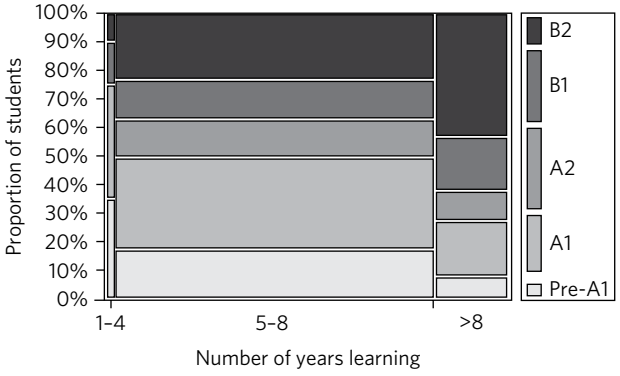
The main survey in Croatia was conducted during the school year 2010–11, from 1 March until 28 March, in 144 sampled schools. English was tested in 68 schools, while German was tested in 69 schools. Since both languages were tested in seven schools, there were 151 successful test administrations, and a total of 3,342 students participated.

ESLC was conducted in primary schools on a representative sample of eighth grade students (ages 13 and 14). A total of 1,109 students were tested in English (49.6%). These students had been learning English for different periods of time: According to the data for primary schools in the school year 2010–11, of the 472,250 students enrolled in primary schools, 10.6% learned English from the first grade. Croatia is one of the participating countries where children start learning English from the first grade or before (European Commission 2012:32). Most students in Croatian schools learn English as their first foreign language and German as their second. Croatian students at ISCED 1 (beginning of primary school) have two first foreign language lessons per week, i.e. at least 70 lessons per year. One lesson lasts 45 minutes. As for students at ISCED 2 (lower-secondary school), they have three first foreign language lessons per week, or at least

105 lessons per year. In the ESLC Report, teaching time is calculated as 60-minute periods and the values are rounded to the nearest integer. According to this calculation, students in Croatian schools have two first foreign language lessons per week. In addition to having the possibility of choosing a foreign language as an optional subject, students in Croatia can participate in extra foreign language lessons (both enrichment and remedial lessons). However, in just a few schools students have the option of learning a second foreign language from the first grade or an ancient language from the fifth grade (Latin) or seventh grade (Greek). The results of the ESLC show that more foreign languages on offer in schools has a positive effect on test results (European Commission 2012:56), and that schools in Croatia 'offer on average only slightly more than two foreign languages (a mean of less than 2.5)' (European Commission 2012:56). In comparison with other participating adjudicated entities[1], this is a very low mean.

According to the data obtained from the ESLC, the majority of participating Croatian students (80%) had been learning English for five to eight years, while approximately 18% of students had been learning English since kindergarten, i.e. for more than eight years. Figure 1 shows that the aforementioned 18% of students are more successful at English reading than the majority of Croatian students. In total, 43% of those students achieve B2 and 19% achieve B1. If we compare these results with the results of the 80% of students who had been learning English for a shorter period of time, from five to eight years, the young learners outperform them by 20% at B2 and 5% at B1.

Figure 1: Results in English reading by number of years learning



The results of the ESLC, presented in Figure 2, indicate that students who started learning the foreign language at an early age are also more successful at English listening. According to the ESLC results, 51% of Croatian students achieve B2 in listening, and 22% achieve B1. If we compare the achievements of students who had been learning English since kindergarten (>8) with the achievements of students who started learning English in primary school (5–8) the biggest difference is visible at pre-A1 level, where the number of students who started learning English at an earlier age is two times smaller than the number of students who started learning their first foreign language in primary school. Furthermore, we can point out the difference between the two categories of students at B2: approximately 20% more students who started learning English at an earlier age achieved this level.

Figure 2: Results in English listening by number of years learning

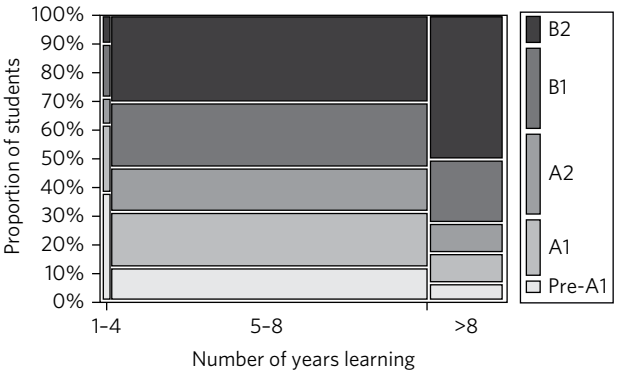
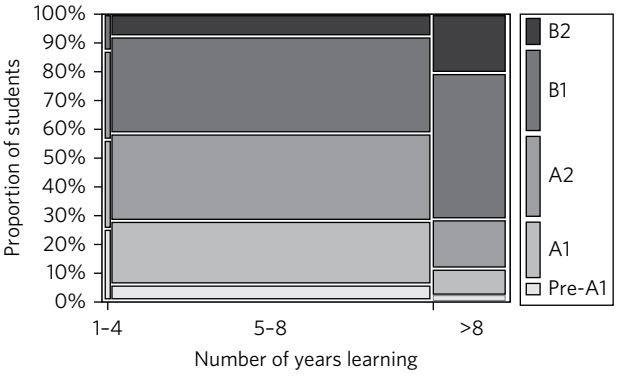


Figure 3: Results in English writing by number of years learning



When it came to writing skills, students participating in the ESLC needed to demonstrate productive foreign language proficiency. As expected, the results for English writing, presented in Figure 3, are somewhat lower in comparison with the results in reading and listening. As in previous analyses, here we also see that students who started learning foreign languages at an earlier age are more successful at all levels of writing. According to the results of the ESLC, the effects of the onset of foreign language teaching on first target language writing are mostly significant, even more so for writing than for reading and listening (European Commission 2012:55). This means that students who started learning English at an earlier age achieve higher results in first target language writing than other students. The fact that young learners have more developed writing skills is interesting since there is more evidence that early foreign language learning influences oral skills and pronunciation. However, the achievement demonstrated by young learners probably stems from longer exposure to the structures of the language in question, since

[1] Countries or communities participating in the ESLC project.

children who attend kindergarten or preschool generally have limited writing skills.

Regarding the other skills, in the category of students learning foreign languages since the first grade of primary school, we note that the number of students at pre-A1 and B2 levels is lower and they are more evenly distributed in the remaining three categories. Twenty-three per cent of students achieve A1, 31% achieve A2 and 34% achieve B1. If we take a look at the results of the students who had been learning foreign languages for more than eight years, 88% of them achieve A2 and higher. Only 5% of students are at pre-A1 and 20% are at A1.

If we compare the results of Croatian students with the results of students in countries that delay the onset of compulsory foreign language education until fifth grade (the French and Flemish Community of Belgium, Bulgaria and Netherlands), it is evident that students from those countries achieve lower results. For instance, if we look at CEFR levels achieved in English reading and listening, the performance of students in Bulgaria and the French Community of Belgium is lower than the performance of Croatian students. However, early foreign language education may not be the only factor influencing the overall lower results of students in the aforementioned countries. Namely, students in the Netherlands and in the Flemish Community of Belgium are among the best in all English skills, and they start learning English at a later age (European Commission 2012:23–24).

Conclusion

The results from the ESLC are extremely significant for the Croatian educational system as a whole. The data collected by the background questionnaires includes information about teaching methods, attitudes, students' work habits, motivation for foreign language learning and other factors that may have affected the final results.

Until the implementation of this survey, the Republic of Croatia has never had such an insight into foreign language teaching and instructional methods. Therefore, the results are invaluable for the further development of foreign language teaching and the creation of educational policies in Croatia.

Thanks to the ESLC, we gained a comprehensive view of the status of foreign languages in Croatia and the position of Croatian students amongst their European counterparts. Although our students are the youngest and due to methodological issues cannot be adequately compared with students in other adjudicated entities, we are satisfied with their achievement. However, there is always room for improvement, especially regarding the second target language, in which the results can and should be better.

On the basis of the results of the ESLC, Croatia will define future policies regarding the improvement of foreign language teaching. The Ministry of Science, Education and Sports has recently made preschool education compulsory for all children in Croatia, which leaves open the possibility of introducing compulsory foreign language education at an even earlier age. Results of the ESLC seem to corroborate the assumption that young learners usually become more proficient in foreign languages. Therefore, it is important for policy makers to recognise that an early start is beneficial and that it has a positive effect on test results. However, older students may achieve equal levels of proficiency if they are motivated enough and exposed to the foreign language on a daily basis. More foreign languages on offer in schools may also contribute to the development of linguistic proficiency in languages other than English. Along with the implementation and improvement of early language learning programmes in preschool institutions, it would be useful to promote bilingual communication in the home environment and thus motivate young learners to use the foreign language in their free time, developing a positive attitude towards foreign languages in general. However, any language learning programme depends on sufficient funding and support from the authorities, as well as quality assurance. Only in this way will language learning programmes (early and otherwise) in Croatia reach their full potential.

References and further reading

Bartolović, B (1993) Young children's cognitive abilities in learning foreign languages, in Vilke, M and Vrhovac, Y (Eds) *Children and Foreign Languages*, Zagreb: University of Zagreb, 27–44.

Clark, B (2002) First- and second-language acquisition in early childhood, in Rothenberg, D (Ed.), *Issues in Early Childhood Education: Curriculum, Teacher Education, & Dissemination of Information (Proceedings of the Lilian Katz Symposium, November 5–7, 2000)*, Urbana-Champaign: University of Illinois, 181–188.

Domínguez, R and Pessoa, S (2005) Early versus late start in foreign language education: documenting achievements, *Foreign Language Annals* 38 (4), 473–483.

Dörnyei, Z (1998) Motivation in second and foreign language learning, *Language Teaching* 31 (3), 117–135, available online: http://www.zoltandornyei.co.uk/uploads/1998-dornyei-lt.pdf

Enever, J (2009) Why applied linguistics is not enough. Contemporary early foreign language learning policy in Europe: a critical analysis, *BAAL Proceedings* 37–38.

European Commission (2012) *First European Survey on Language Competences: Final Report*, Luxembourg: Publications Office of the European Union, available online: ec.europa.eu/languages/eslc/index.html

Griva, E and Sivropoulou E (2009) Implementation and evaluation of an early foreign language project in kindergarten, *The Early Childhood Education Journal* 37 (1), 79–87, available online: dx.doi.org/10.1007/s10643-009-0314-3

Marinova-Todd, S, Marshall, D and Snow, C (2000) Three misconceptions about age and L2 learning, *TESOL Quarterly* 34 (1), 9–34.

Mihaljevic Djigunovic, J (1995) Attitudes of young foreign language learners: A follow-up study, in Vilke, M and Vrhovac, Y (Eds) *Children and Foreign Languages*, Zagreb: University of Zagreb, 17–33.

Ministry of Science, Education and Sports (2006) *Croatian National Educational Standard: Syllabus for the English Language*, Zagreb, available online: public.mzos.hr/Default.aspx?sec=2197

Nikolov, M and Curtain, H (Eds) (2000) *An Early Start: Young Learners and Modern Languages in Europe and Beyond*, Council of Europe: European Centre for Modern Languages, available online: www.poliglotti4.eu/docs/Research/An_Early_Start_Young_Learners_and_Modern_Languages_in_Europe_and_Beyond.pdf

Szpotowicz, M, Mihaljevic Djigunovic, J and Enever, J (2009) Early language learning in Europe: a multinational, longitudinal study, in Enever, J, Moon, J and Raman, U (Eds) *Young Learner English Language Policy and Implementation: International Perspectives*, Reading: Garnet Publishing, 147–153.

Vilke, M (1993) Early foreign language teaching in Croatian primary schools, in Vilke, M and Vrhovac, Y (Eds) *Children and Foreign Languages*, Zagreb: University of Zagreb, 10–26.

Vilke, M (1995) Children and foreign languages in Croatian primary schools, in Vilke, M and Vrhovac, Y (Eds) *Children and Foreign Languages*, Zagreb: University of Zagreb, 1–15.

Acknowledgements

As the National Research Coordinator for Croatia, I am grateful to the Government of the Republic of Croatia and the Ministry of Science, Education and Sports for recognising the opportunity to participate in this important survey on assessment of language competences of European students and for the financial support provided for the administration of the survey.

Special thanks to Nataša Tepić for creating the box plots for the purpose of this article, and Maja Kušan for the translation.

Reflections on the European Survey on Language Competences: Looking back, looking forwards

KAREN ASHTON INSTITUTE OF EDUCATION, MASSEY UNIVERSITY, NEW ZEALAND

Introduction

The European Survey on Language Competences (ESLC), the first survey or research project of its kind, has provided empirical evidence that did not exist before. For the first time, there is now hard data on outcomes of language education in and across a range of European countries as well as indicators of what makes for successful language learning or successful language learners. The accomplishments of the ESLC owe much to the formation of positive relationships, in particular the close collaboration across the 16 educational systems[1] that participated in the ESLC, the European Commission and the eight international partners led by Cambridge English Language Assessment under a consortium called 'SurveyLang', who were contracted by the European Commission to carry out the survey. This short article takes the opportunity to reflect back on some of the challenges, limitations and lessons learned during the busy 4-year project with a forward focus on the next round of the ESLC.

The Terms of Reference

The Terms of Reference provided by the European Commission (2007) were agreed by the European Indicator of Language Competence (EILC) Advisory Board, which was set up late 2006 in preparation for the start of the ESLC (February 2008). The Advisory Board was made up of national expert representatives from each European Union member state, including those that chose not to participate in the first round of the ESLC. Their tasks were to advise the European Commission in establishing the technical parameters for the survey, i.e. the Terms of Reference, as well as to comment on the progress made by SurveyLang and any technical aspects throughout the 4–year period.

The Terms of Reference set out the overall objectives of the ESLC, e.g. the collection of data to enable comparisons between countries, as well as practical aspects such as the timing of the survey and specific reporting objectives to be met by SurveyLang. The most important elements of the Terms of Reference, however, were the technical parameters (or the 'framework') set out for the survey. Several aspects which impacted most significantly on the design of the survey are discussed below, together with some thoughts and reflections.

Skills and levels tested

The Terms of Reference state that only the skills of listening, reading and writing were to be tested. The testing of speaking was deemed too logistically complex to administer in the first round of the ESLC. Part of the reasoning behind this was the expectation that the majority of countries would embrace computer-based (CB) testing, which was not far enough advanced to cope with CB testing of speaking on a large scale, but additionally and related to this, there was also the acknowledgement that things should be kept as 'simple' as possible. There had been some initial scepticism from members of the Advisory Board about the feasibility of the survey and the European Commission recognised that minimising potential obstacles was important in helping to ensure the success of the survey. It should be noted that the views of the Advisory Board became much more positive early on in the project and remained that way throughout with members congratulating SurveyLang on achieving what they previously considered unachievable.

The Terms of Reference were clear that outcomes for the three skills tested should be reported (and thus compared across countries) in terms of Levels A1 to B2 of the Common European Framework of Reference for Languages (CEFR)

[1] Fourteen countries participated but they represented 16 educational systems as Belgium's three linguistic communities participated separately. For ease of reading, this article uses the term *country* to denote educational system.

(Council of Europe 2001). The dependent variable measured by the survey is therefore the language ability of secondary school students. To derive a measure comparable across languages this was defined as the ability to use language purposefully, either to understand spoken or written texts or to express oneself in writing. The omission of speaking may impact on the interpretation of the results. Although the Advisory Board deemed it necessary for reporting to create a global average across the three skills, and although the results have in some cases reinforced what was already felt, e.g. country x is good at languages, and country y is bad at languages, the impact of the decision to exclude speaking is unknown and cannot be assessed. The global averages might have looked somewhat different had the skill of speaking been tested. The European Commission has stated that speaking will be included in the second round of the ESLC and this is an important step forward in further understanding the ability of students to use language purposefully.

Tested languages

The Terms of Reference specified that the two most taught languages out of English, French, German, Italian and Spanish be tested in each country participating in the survey. These languages are the five most taught European languages across Europe. Several issues arose here. First, some countries were not pleased that their national language was not included among the five listed. Second, the limited list of languages available for testing meant that in several cases the first and second most taught languages were not tested. For example, in the French-speaking Belgium community, Dutch is the first most taught language and English the second. This meant that they had to test their second (English) and third (German) most taught languages. Similarly, in Bulgaria, as Russian is the second most taught language, Bulgaria had to test its first (English) and third (German) most taught languages. This is an important caveat in interpreting results and in making comparisons across countries.

Additionally, among the five languages available for testing, as may be expected, English dominated. Fifteen out of the 16 countries tested in English, mostly as the first language. While French was tested in six countries, and German in eight, Spanish was only tested in two countries while Italian was only tested in Malta. Clearly the results of the survey tell us a lot more about the learning and teaching of English across Europe than they do for Spanish or Italian.

An interesting situation arose where one country wanted to test in more than two of the five languages. Although SurveyLang would very much liked to have offered this, in practice, the complexities of needing two independent samples for each tested language meant that it would not have been possible to draw further independent samples of the size required to test additional languages.

Given that the European Commission would like to increase the number of languages tested in the second round of the ESLC (there has been mention of testing the official languages of all EU member states) different methods for data collection may need to be considered. This point will be returned to at the conclusion of this article.

Tested level

The Terms of Reference specify the final year of lower secondary education (ISCED 2) as the main testing grade. However, as not all countries offer all languages at this level, the second year of upper secondary (ISCED 3) was specified as an additional testing grade. Students were only eligible for testing if they had received a minimum of one year's formal language tuition in the tested language. The Terms of Reference note two important issues related to this decision. First, that the 'age and time during which pupils have been learning a foreign language will be different' across countries, and second, that language learning is voluntary in some contexts but compulsory in others (European Commission 2007:8). The context of learning languages is very different from the context of subjects tested by other international surveys such as the Programme for International Student Assessment (PISA), the Third International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS), which are compulsory for all students, making comparisons across countries easier.

Beyond the technical aspects outlined in the Terms of Reference, there were other elements worth briefly reflecting on. These are detailed below.

Country participation

The survey was open to 'the 27 countries in the European Union, candidate countries that also take part in the Lifelong Learning Programme (Turkey, Croatia) and the [European Economic Area] EEA Member States (Iceland, Lichtenstein and Norway)' (European Commission 2007:14). Countries therefore had the choice as to whether they participated or not. As with other international surveys, countries paid for the costs required for managing the survey in-country; however, unlike other surveys, they did not pay a participation fee. A specific participation fee was not necessary as the survey was funded by the Lifelong Learning Programme fund, which the above countries contribute to.

The lack of a participation fee was of course positive but it also had drawbacks. It meant that the number of countries participating was not clear from the outset. Although the project started in February 2008, final participation was not clear until late 2009, a few months before the Field Trial in February/March 2010. In the end 16 educational systems (albeit 14 countries as Belgium's three linguistic communities participated separately) participated. Reasons for not participating ranged from the financial implications (given the serious financial uncertainty across Europe), concerns about where countries would rank, the desire to be spectators in the first round to see what happened, concerns about survey 'fatigue' (particularly for small countries already committed to participating in other international surveys) and, for one country, principled objections to indicators. In this sense, this first round acts as a kind of pilot for future rounds. It was encouraging that many non-participating countries regularly attended and participated in Advisory Board meetings and maintained a strong interest in the progress of the survey. In order for the survey to become a truly 'European' survey, it is hoped that the success of the first round contributes to greater participation in the second round.

Administration challenges

The administration of the survey was very complex. The ESLC had a targeted design with overlapping difficulty levels, e.g. A1/A2, A2/B1 and B1/B2 (see Robinson's article in this issue for more details). This targeted approach allowed for the collection of more valid data and meant that students sat tests at an appropriate level of difficulty and challenge. To add to this complexity each student was tested in only two out of three skills in order to limit the testing time for students. This kind of matrix design was possible as results and comparisons were analysed at the level of schools and country rather than sampled student. However, given that students were sampled randomly from each school rather than sampled as a whole class, the room planning, timetabling and logistics were very complex. For example, at any given school it was possible for all three levels to be tested. If the school had opted for paper-based (PB) testing, different administrations were needed for the listening tests. In this aspect, the ESLC is very different from other international surveys where all students receive the same test. To ensure that the correct students sat the correct tests, SurveyLang provided countries with DVDs with personalised student tests (using IDs rather than names) in pdf format. It also developed room plans for every possible combination of testing and room availability, 32 plans in total. As one can imagine, this was not particularly popular with administrators and is something that would need further consideration in the second round, particularly if speaking is to be tested as well. In addition to reviewing the design decisions, wider use of CB testing would limit these challenges as students could take different listening tests within the same room.

Computer-based (CB) testing/ Paper-based (PB) testing

CB testing was seen as 'the optimal step forward in relation to the survey' (European Commission 2007:3) and it was hoped that CB testing would be widely embraced. However, the reality was very different and countries' preparedness and enthusiasm for CB testing was overestimated. As logistically and operationally it was too complex to offer both modes within a single school, SurveyLang set the decision at school level; however, most frequently the decision was taken at country level, with the majority of countries opting for PB testing. Nine countries opted for full PB, four for full CB, and three used a combination, although even in these three countries PB dominated. Reasons for deciding at country level to use PB rather than CB testing ranged from bad experiences in other surveys, negative perceptions about the difficulty and reliability of CB testing, concerns that school computers did not meet the required specification, and that technical staff at school would find the task difficult. The preparation work for CB testing, supported by SurveyLang, was more involved than for PB; however, the room planning logistics were a lot less complex, printing costs were lower, and data entry requirements considerably lower. Overall, very few technical issues were experienced making CB testing less time consuming and costly than PB testing. What is clear from the first round of the survey is that buy-in for CB testing cannot be assumed. Concerted efforts to promote CB testing should be considered for the second round.

The way forwards?

Key considerations outlined here include the need for speaking to be tested in the second round to ensure a more complete assessment of language proficiency. This is a substantive task in itself, particularly if CB testing of speaking is to be widely used. To ensure greater uptake of the CB format, additional 'marketing' and promotion is needed. This needs to be a concerted effort; the benefits of CB testing cannot simply be assumed. Factors for country non-participation also need to be addressed. The results and success from this survey should be used to promote and increase participation in the second round. Although nothing can be done about external factors such as the financial crisis, collaboration or agreement between survey providers could avoid the survey 'fatigue' reported by countries.

More important is the need to recognise that the second round of the ESLC has several aims. First, it needs to benchmark student performance so that the results of the second round are comparable to the first. Policy makers will want to know, for example, whether there have been any changes in the proficiency levels of students in particular countries and how countries are doing relative to each other. This will require another large-scale survey administration with formal sampling procedures and would be a 'compulsory' element of participating in the second round. However, what is clear from the above discussion is that survey methodology needs to be complemented with other methods so that a more complete picture and understanding of language learning, teaching and ability across the different learning contexts in Europe, and beyond the dominance of English, can be obtained. One possible area to explore includes helping countries to link school (or other) language exams to the ESLC language tests. This would allow for data to be collected for languages not in the current list of five. Another possibility is the option to administer an additional questionnaire to sampled students (or a sub-set of students) covering policy issues of high relevance to that country in greater depth than is possible in the main survey data collection. Similarly, some countries may be interested in interviews or focus groups to look at a particular issue of interest in more depth. The above suggestions could work as in-country case studies designed to complement the survey data collection within an overall mixed-methods approach. Although there would be less formal sampling, the data would provide additional understanding of the current picture of language learning beyond what is possible in a formal survey. Any such complementary methods should not be imposed on countries through the Terms of Reference. Rather, they should be presented as options that countries as the key stakeholders and end-users of the data can opt into and help refine, together with the contractor and the European Commission. In this way, although the additional data collected will not be systematic across countries, it would help in working towards the second aim of the ESLC, which is to better understand language learning and teaching in Europe.

References

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

European Commission (2007) *Terms of Reference: Tender no. 21 'European Survey on Language Competences'*, Contracting Authority: European Commission.

The European Survey on Language Competences and the media

STEPHEN MCKENNA COMMUNICATIONS AND STAKEHOLDER RELATIONS GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

The results of the first European Survey on Language Competences (ESLC) were announced by the European Commission on 21 June 2012, with simultaneous release of press statements by the European Commission, and by many of the National Research Coordinators or their Ministries of Education. The ESLC Final Report, Technical Report and Executive Summary were released at the same time (available online: ec.europa.eu/languages/eslc/index.html).

The results attracted widespread media coverage in the participating countries. Overwhelmingly, this coverage focused on the 'league table' showing relative performance by country, published in the Executive Summary, and on the proportion of students achieving Level B1 in the participating countries. Particularly extensive coverage was generated in countries ranked low in this table, with headlines often focusing on particular areas of perceived weakness, as in the following examples from national media in Spain and the UK:

> El 63% del los estudiantes que acaban la ESO tienen dificultades para entender el inglés oral [63% of students who complete secondary education have difficulty understanding spoken English] (ABC, 21 June 2012, available online: www.abc.es/20120621/sociedad/abci-estudio-idiomas-linguistico-espaa-201206211656.html)
>
> Pupils in England worst for using languages independently (BBC News online, 21 June 2012, available online: www.bbc.co.uk/news/education-18531751)

This focus on particularly negative outcomes of the survey is in line with a widely observable tendency of news media to focus on 'bad news' when reporting educational stories.

From the point of view of the language testing profession, the coverage of the ESLC results has some striking – if unsurprising – characteristics:

- heavy focus on the 'league table' to the exclusion of other indicators such as the factors which contribute to successful learning, which are presented in detail in the Summary Report and other documentation
- a strong preference for citing the opinions of political officials rather than seeking input from educationalists or academic commentators.

In a number of cases, however, Project Director Dr Neil Jones was able to put forward a professional point of view. Thus, in a letter to *The Times*, Jones is quoted as saying:

> The findings highlight what specialists in language learning have known for a long time – that communication skills need to be at the heart of language teaching. Students need to be taught and encouraged to treat foreign languages as part of their everyday lives (*The Times*, 22 June 2012).

Jones developed this theme in an interview published a month later in a leading national newspaper in France, which was quickly followed by further coverage of the Survey in the French media, including an interview with the Education Minister on national TV news[1].

The reporting of Jones's comments in national newspapers show that it is possible for the voice of the language testing profession to be heard in public discussion and debate, if academics are prepared to engage with the media on its own terms, offering prompt, expert and non-technical commentary. The strength of this approach is also illustrated in articles by Michael Milanovic, Chief Executive of Cambridge English Language Assessment, giving advice on setting appropriate language standards for healthcare professionals, widely published in specialist healthcare publications in the UK (see, for example, the following article: careers.bmj.com/careers/advice/view-article.html?id=20004262).

Above all, the coverage of the ESLC shows clearly both the potential for assessment specialists to engage in discussion of public policy issues and the risk that if language testing experts choose not to do so, discussion and policy will be shaped by less well-informed commentators.

[1] L'élève français, ce cancre en langues étrangères (*Le Monde*, 22 July 2012, available online: www.lemonde.fr/education/article/2012/07/22/l-eleve-francais-ce-cancre-en-langues-etrangeres_1736714_1473685.html [the headline is not taken from Jones's comments]

Examining textual features of reading texts – a practical approach

ANTHONY GREEN CRELLA, UNIVERSITY OF BEDFORDSHIRE, UK
HANAN KHALIFA RESEARCH AND VALIDATION GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT
CYRIL J WEIR POWDRILL CHAIR IN ENGLISH LANGUAGE ACQUISITION, CRELLA, UNIVERSITY OF BEDFORDSHIRE, UK

Introduction

Until recently only very limited quantitative evidence has been available to indicate what distinguishes texts targeting test takers at different levels of proficiency. Traditional checklist-based approaches to text description have proved to be problematic, especially in the context of operational test development, because they are time consuming and have limited reliability. Recent developments in computational linguistics have produced a number of automated tools for text description that could be of value to test developers in supporting consistency and authenticity in text selection. Through the use of such tools, relevant textual features such as those listed by Khalifa and Weir (2009) in their socio-cognitive reading model might be more easily captured and accounted for in operational test development. Texts matching specified characteristics might be more readily identified, enhancing test validity without impacting practicality.

This study aims to investigate and suggest a set of key features for the *a priori* evaluation of texts at the selection stage via the use of computational tools, namely, *Coh-Metrix* (Graesser, McNamara, Louwerse and Cai 2004, McNamara, Louwerse, Cai and Graesser 2005) and *VocabProfile* (Cobb 2003). The study focuses on reading texts at Common European Framework of Reference (CEFR) C1 level as represented in the *Cambridge English: Advanced (CAE)* examination and seeks out text features that most consistently distinguish *Cambridge English: Advanced* level texts from those at adjacent levels, i.e. *Cambridge English: Proficiency (CPE)* set at C2 level and *Cambridge English: First (FCE)* targeting CEFR B2 level. The tools were selected because of practical considerations in terms of software availability, cost, flexibility in handling extended text and the interpretability of output as well as validity considerations in the abovementioned reading model. *Coh-Metrix* was particularly attractive because it is based on cognitive accounts of reading that are compatible with the processing model advocated by Khalifa and Weir (2009). Although both tools can be accessed freely via web interfaces, they do have some serious limitations. These are considered in the literature review below, in the reporting of results and are reflected in the recommendations made for operational test text selection.

Literature review on analysing text complexity

The literature on potential sources of text complexity is extensive and so this brief survey will be necessarily limited. It focuses initially on studies that considered a broad range of textual parameters relevant to this study, and then goes on to overview the potential contribution of automated text analyses, specifically using *Coh-Metrix*, to enhance our knowledge and understanding of text complexity.

Analysing text complexity – subjective approaches

A number of descriptive typologies of text characteristics have been designed for use in test development and validation studies (e.g. Alderson, Figueras, Kuijper, Nold, Takala and Tardieu 2004, Bachman, Davidson, Ryan and Choi 1995, Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt and Schedl 2000, Fortus, Coriat and Fund 1998, Freedle and Kostin 1993, and Khalifa and Weir 2009). There appears to be a measure of consensus in the subjective judgements of these different authors on the features to be addressed when considering text complexity.

Bachman et al's (1995) test comparability study identified textual properties such as the nature of the text, length, vocabulary, grammar, cohesion, distribution of new information, type of information, topic, genre, rhetorical organisation and illocutionary acts. Freedle and Kostin (1993) took into consideration referentials, rhetorical organisers, fronted structures, vocabulary, concreteness/abstractness, subject matter, coherence, length of various segments such as word, sentence, paragraphs, and passage as text-related variables. Fortus et al (1998) investigated length, number of negations, number of referential markers, vocabulary, grammatical complexity, abstractness, topic and rhetorical structure as textual variables contributing to the level of difficulty of reading comprehension items.

Enright et al (2000) identified three groups of salient textual features: grammatical/discourse features, pragmatic/rhetorical features and linguistic variables. Alderson et al (2004) included text source, authenticity, discourse type, domain, topic, nature of content, text length, vocabulary and grammar as relevant features for text analysis. Khalifa and Weir (2009) examined the contextual features proposed in the research literature and established a subset which enabled Cambridge English Language Assessment to make criterial distinctions between levels of proficiency in their reading examinations.

Analysing text complexity – an automated solution

Recent advances in automated textual analysis and computational linguistics have now made it feasible to provide more quantitative approaches focusing analytically on a wide range of individual characteristics of texts (Crossley, Greenfield and McNamara 2008, Crossley, Louwerse and McNamara 2008, Graesser et al 2004, Green

2012). New technologies now offer examination boards the potential for a more systematic, efficient way of describing a number of the contextual parameters in texts (see Green, Ünaldi and Weir 2010) and to compare automatically and accurately the various contextual characteristics of the range of written texts at different levels of ability. The following provides a brief description of one such technological solution, i.e. *Coh-Metrix*.

Coh-Metrix is a computational tool which incorporates measures of lexical complexity, structural complexity and a text-level representation to enable a fine-grained analysis of a text. McNamara, Graesser and Louwerse (in press) describe the capabilities of *Coh-Metrix* as follows:

> Coh-Metrix is a tool that provides numerous indices of language automatically. [...] This tool augments conventional readability formulas, such as the Flesch-Kincaid and Flesch Reading Ease, with computational indices of text cohesion as well as an assortment of characteristics of words, sentences, and discourse. Coh-Metrix uses lexicons, part-of speech classifiers, syntactic parsers, latent semantic analysis (a statistical representation of world knowledge based on corpus analyses), and several other components that are widely used in linguistics.

Graesser, McNamara and Kulikowich argue that such 'automated analysis is unquestionably more reliable and objective than approaches that involve humans annotating and rating texts by hand' (2011:223–234). In addition, we might speculate that automated analysis brings significant benefits in terms of its labour-saving efficiency.

At the time of this study, the publicly available, online version of *Coh-Metrix* offered a total of 54 separate textual indices which were selected by the software developers from a much larger repository of over 600 indices, and categorised as follows (the bracketed value is the number of indices within the category):

i) *readability indices* (measures of reading difficulty) – [2]
ii) *general word and text information* (basic counts, frequencies, concreteness, hypernymy) – [14]
iii) *syntactic indices* (complexity, composition and frequency of classes/constituents) – [22]
iv) *referential and semantic indices* (anaphor, co-reference, latent semantic analysis (LSA)) – [10]
v) *situational model dimensions* (causal, intentional, temporal, spatial dimensions) – [6].

For ease of reference, application and interpretation, these five categories can be further collapsed to form three broad strands relating to *lexical complexity* (i and ii), *syntactic complexity* (iii), and *text-level representation* (iv and v). These overarching strands are frequently referred to by the developers in reports of their *Coh-Metrix* text analyses. The strands also reflect the widely accepted (if more intuitive) trinitarian approach of analysing text at the *word*, *sentence* and *discourse* level. More detail is provided below on each of these three strands of analysis based upon the available *Coh-Metrix* literature. For reasons of brevity, not all of the available indices are described below; instead, selected indices from within each strand are presented and discussed to provide some level of illustration and explanation of how *Coh-Metrix* functions. These selected indices are also some of those which past studies suggest are likely to be most productive or informative in an analysis of text complexity using *Coh-Metrix*.

Lexical complexity strand and selected indices

*Number of words (*CohMx35*)*

Potentially more important than simple text length is the density and complexity of idea units within a text (Bachman 1990). Unfortunately, *Coh-Metrix* cannot analyse this feature and it thus remains a task for human analysts to undertake. Nonetheless, in a testing context, the longer the written output, the more complex the text, the more likely it is to contain a greater number of idea units. The literature makes it clear that *Coh-Metrix* works best when used with texts of at least 200 words, particularly for indices such as type-token ratio. It seems that applying *Coh-Metrix* to texts below 200 words in length risks resulting in a less accurate picture of a textual complexity.

*Readability formulas: Flesch Reading Ease Score (*CohMx39*) and Flesch-Kincaid Grade Level (*CohMx40*)*

In test development, readability formulas (e.g. Flesch Reading Ease/Flesch-Kincaid Grade Level) combining relatively basic syntactic and lexical features such as word and sentence length are often used as convenient indicators of text complexity. However, these have been criticised as inadequate for revealing textual complexity (see, for example, Gervasi and Ambriola 2002 and Masi 2002), for ignoring 'dozens of language and discourse components that are theoretically expected to influence comprehension difficulty' (Graesser et al 2004:194) and as being inappropriate for L2 readers (Brown 1997). Graesser et al (2011:224) explain that *Coh-Metrix* was specifically developed to analyse texts on multiple characteristics and levels of discourse, including an automated metric of text cohesion (hence the label *Coh-Metrix*).

Syntactic complexity strand and selected indices

Crossley, Greenfield et al (2008:482) observe that:

> A reading text is processed linearly, with the reader decoding it word by word; but, as he or she reads, the reader also has to assemble decoded items into a larger scale syntactic structure. Clearly, the cognitive demands imposed by this operation vary considerably according to how complex the structure is (Perfetti, Landi, and Oakhill 2005).

Texts with less complex grammar tend on the whole to be easier than texts with more complex grammar. A considerable number of indices have been suggested in the literature for the estimation of grammatical complexity (see Ortega 2003, Wolfe-Quintero, Inagaki and Kim 1998). Some of the quantitative measures available through *Coh-Metrix* and some of the seemingly most productive indices are reviewed below.

*Mean number of modifiers per noun phrase (*CohMx41*)*

The inclusion of modifiers increases the length and complexity of the string of words that a reader has to hold in the mind while imposing a syntactic pattern upon it. The mean number of modifiers per noun phrase (NP) is an index of the complexity of referencing expressions. Barker (1998) argues that noun phrases carry much of the information in a text and computerised systems that attempt to acquire knowledge from text must first decompose complex noun phrases to get access to that information. Graesser et al (2004) suggest that

sentences with difficult syntactic composition have a higher ratio of constituents per noun phrase than do sentences with simple syntax.

*Mean number of words before the main verb of a main clause (*CohMx43*)*

This refers to the number of words that appear before the main verb of the main clause in the sentences of a text. According to the guidance provided for users of *Coh-Metrix*, sentences that have many words before the main verb are taxing on working memory and thus difficult to process. However, their justification is not altogether convincing. The authors refer to working memory as a very general notion and do not specify how it operates in this case. The explanation may relate to parsing. The words that occur before the verb are the first in a sentence to be analysed, and the longer the subject NP, the greater will be the burden imposed at this early stage on working memory.

Text-level representation strand and selected indices

In addition to automated lexical and syntactic analysis, *Coh-Metrix* analyses a variety of indices that move beyond the word and sentence level, and which relate to the discourse level of the text. Some of the key indices of interest relate to features of cohesion and coherence, e.g. *CohMx* 16, 18, 21, 26, 58 and 60 (see the section *The 17 Differentiating Indices – Key Finding 1* for a detailed discussion of these).

Cohesive devices cue the reader on how to form a coherent representation. The coherence relations are constructed in the mind of the reader and depend on the skills and knowledge that the reader brings to the situation. In other words, coherence is a psychological construct, whereas cohesion is a textual construct. It has been argued that explicit cohesive devices can help in establishing textual coherence (Goldman and Rakestraw 2000), though Alderson (2000) notes that an absence of cohesive devices does not seriously damage comprehension when the topic is relatively familiar to readers.

McNamara et al (in press) provide a useful theoretical discussion of some of the key sources of text difficulty linked to cohesion, highlighting co-reference (argument overlap and stem overlap), verb cohesion, connectives and causal cohesion as among the key indices. The McNamara et al (in press) paper also reports on an empirical study to investigate the role of cohesion in text difficulty across a corpus of texts varying in grade level and genre (narrative/expository). Findings suggested that referential cohesion (repetition of nouns or verbs) tended to increase across grade levels, though lower grade texts showed lower repetition of object nouns and higher verb cohesion. The researchers interpret this to suggest that the less overlap in objects is compensated by more overlap in actions, indicating a trade-off between difficulty at the lexical and cohesion levels.

Summary

This brief literature review surveys some of the growing body of research reporting on the use of *Coh-Metrix* to provide multi-level analyses of text characteristics. Recent research by Graesser et al (2011) highlights five major factors which they believe can account for most of the variance in texts across grade levels and text categories: word concreteness, syntactic simplicity, referential cohesion, causal cohesion and narrativity. Outcomes from previous studies using *Coh-Metrix* would seem to provide sound theoretical and empirical justification for applying the tool to the corpus of reading texts gathered for this study, and suggest some specific avenues for investigation.

Application of computational tools to B2, C1 and C2 test material

Having reviewed the literature on text complexity and identified suitable tools for text analysis, this paper now focuses on applying these tools to *Cambridge English: First* (B2), *Cambridge English: Advanced* (C1) and *Cambridge English: Proficiency* (C2) to identify features that changed by level and so might be closely identified with perceptions of text difficulty. The study was restricted to text characteristics that could be measured *automatically* through freely available software.

A total of 166 reading texts/extracts were used in the analysis, i.e. 48 for *Cambridge English: First*, 49 for *Cambridge English: Advanced* and 69 for *Cambridge English: Proficiency*. The texts were transformed to ASCII test files and analysed via *Coh-Metrix* and *VocabProfile*. The output of these analyses informed the subsequent process of identifying indices that effectively distinguished the levels of text difficulty currently operationalised in the Cambridge English examinations. Independent-Samples Kruskal-Wallis Tests were then performed to test whether the means observed for each textual index were similar across the three levels: *Cambridge English: First*, *Cambridge English: Advanced* and *Cambridge English: Proficiency*. Indices that did not yield any statistically significant difference were eliminated. Results were also plotted against test level and a focus group – a total of six applied linguistics experts identified those that appeared most clearly to differentiate between levels. The focus group considered each feature in relation to the socio-cognitive validity framework (Khalifa and Weir 2009).

Through these processes, a shortlist of 17 candidate indices were identified (from an initial set of 54 *Coh-Metrix* and six *VocabProfile* indices) that appeared to play an important part in determining text difficulty and so might inform operational text selection. Stepwise multiple regression was then employed to determine which of these indices was most criterial in assigning texts to level. The procedures employed to reach a decision on the most useful quantitative indices are described below in detail.

Procedure 1: Descriptive statistics

Descriptive statistics were calculated for all indices for the texts at each of the three levels of proficiency (Table 1).

On the basis of the descriptive statistics, we identified indices which did not exhibit any observable differences across the three proficiency levels: *Cambridge English: First*, *Cambridge English: Advanced* and *Cambridge English: Proficiency*. These indices were considered unlikely to represent useful features for text selection and so were designated for rejection from further analysis. The 13 indices flagged for rejection are: 8, 13, 19, 24, 31, 36, 45, 49, 50, 51, 52, 54 and 59. The numbers used correspond to the coding of the indices in *Coh-Metrix* (see Table 1). We later confirmed through the focus group discussion (see Procedure 4) that none of these

Table 1: Descriptive statistics: All indices

Coh-Metrix indices	Cambridge English: First (n = 48)				Cambridge English: Advanced (n = 49)				Cambridge English: Proficiency (n = 69)			
	Min	Max	Mean	Std. Dev	Min	Max	Mean	Std. Dev	Min	Max	Mean	Std. Dev
7 Incidence of causal verbs, links and particles	40.71	82.11	57.31	10.12	22.47	78.60	53.54	11.29	4.44	67.96	31.02	12.45
8 Ratio of causal particles (cp) to causal verbs (cp divided by cv+1)	0.40	2.00	0.93	0.38	0.10	1.70	0.89	0.36	0.04	3.17	1.06	0.64
9 Incidence of positive additive connectives	13.20	53.25	28.79	7.96	15.36	59.93	30.38	9.58	0.00	41.15	13.04	10.40
10 Incidence of positive temporal connectives	1.88	23.56	12.47	5.73	0.00	17.48	7.88	4.48	0.00	48.10	19.78	11.92
11 Incidence of positive causal connectives	10.07	47.11	25.94	7.68	4.90	45.97	24.29	8.95	0.00	42.80	12.88	7.32
12 Incidence of negative additive connectives	2.93	23.32	11.02	5.14	4.17	30.12	12.39	5.20	0.00	4.90	0.45	1.09
13 Incidence of negative temporal connectives	0.00	3.00	0.50	0.97	0.00	4.00	0.53	1.00	0.00	10.66	1.15	2.21
14 Incidence of negative causal connectives	0.00	5.00	1.48	1.52	0.00	4.00	0.98	1.23	0.10	120.99	57.61	39.64
15 Incidence of all connectives	50.00	106.49	78.11	13.98	44.12	106.62	75.25	14.82	0.07	108.95	20.50	33.26
16 Argument overlap, adjacent, unweighted	0.27	0.70	0.50	0.12	0.10	0.88	0.45	0.16	0.00	1.00	0.38	0.24
17 Stem overlap, adjacent, unweighted	0.05	0.52	0.26	0.13	0.00	0.79	0.32	0.17	29.75	84.44	51.03	11.10
18 Anaphor reference, adjacent, unweighted	0.17	0.84	0.53	0.14	0.11	0.79	0.41	0.16	0.00	0.75	0.37	0.20
19 Argument overlap, all distances	0.22	0.61	0.42	0.11	0.20	0.72	0.40	0.13	0.04	1.00	0.41	0.18
20 Stem overlap, all distances, unweighted	0.05	0.42	0.22	0.10	0.07	0.67	0.29	0.14	0.04	1.00	0.34	0.20
21 Anaphor reference, all distances	0.11	0.55	0.30	0.12	0.03	0.55	0.20	0.12	0.00	0.56	0.18	0.13
22 Noun phrase incidence score (per thousand words)	252.83	319.00	285.25	17.28	234.49	305.79	270.83	18.21	206.90	339.09	265.83	24.28
23 Ratio of pronouns to noun phrases	0.15	0.52	0.34	0.09	0.12	0.48	0.26	0.11	0.03	0.50	0.23	0.12
24 Number of conditional expressions, incidence score	0.00	10.00	2.63	2.49	0.00	7.00	2.35	2.30	0.00	10.71	2.96	2.92
25 Number of negations, incidence score	0.00	16.39	5.83	3.70	0.00	19.92	6.44	4.72	0.00	26.13	8.90	6.25
26 Logical operator incidence score	18.77	53.25	35.72	8.31	16.71	66.27	38.58	11.42	9.80	77.82	44.87	15.22
27 LSA, sentence to sentence adjacent mean	0.07	0.20	0.13	0.03	0.07	0.32	0.15	0.05	0.06	0.43	0.18	0.07
28 LSA, sentences all combinations mean	0.07	0.20	0.13	0.03	0.05	0.28	0.14	0.06	0.06	0.45	0.17	0.07
29 LSA, paragraph to paragraph, mean	0.04	0.47	0.27	0.11	0.00	0.59	0.34	0.12	−0.02	0.66	0.35	0.17
30 Personal pronoun incidence score	39.92	165.69	98.09	29.47	29.22	140.75	70.88	32.36	7.77	144.87	61.78	36.79
31 Mean hypernym values of nouns	4.39	5.76	5.05	0.32	4.23	5.76	4.94	0.36	3.81	5.83	4.86	0.44
32 Mean hypernym values of verbs	1.29	1.79	1.54	0.12	1.06	1.83	1.50	0.16	1.20	1.93	1.58	0.15
33 Number of paragraphs	4.00	21.00	9.35	4.03	1.00	18.00	8.43	4.53	1.00	28.00	4.91	4.36
34 Number of sentences	19.00	58.00	35.08	7.04	9.00	74.00	36.61	16.44	5.00	62.00	19.12	10.00
35 Number of words	476.00	764.00	637.67	63.28	166.00	1285.00	713.33	269.44	179.00	849.00	405.71	152.95
36 Average sentences per paragraph	1.43	8.60	4.29	1.67	2.17	9.17	4.85	1.68	1.18	15.00	4.78	2.38
37 Average words per sentence	11.57	26.05	18.68	3.03	13.97	29.08	20.42	4.01	11.29	49.38	23.39	7.25
38 Average syllables per word	1.30	1.58	1.42	0.07	1.28	1.76	1.54	0.11	1.30	1.93	1.55	0.13
39 Flesch Reading Ease Score (0–100)	51.64	80.61	68.01	7.87	29.87	82.22	56.09	12.05	16.83	83.27	51.64	16.05

Table 1: (continued)

Coh-Metrix indices	Cambridge English: First (n = 48)				Cambridge English: Advanced (n = 49)				Cambridge English: Proficiency (n = 69)			
	Min	Max	Mean	Std. Dev	Min	Max	Mean	Std. Dev	Min	Max	Mean	Std. Dev
40 Flesch-Kincaid Grade Level (0–12)	5.06	12.00	8.41	1.71	5.71	12.00	10.01	1.80	4.70	12.00	10.38	2.03
41 Mean number of modifiers per noun phrase	0.42	1.13	0.75	0.16	0.54	1.37	0.92	0.20	0.55	1.28	0.90	0.16
42 Higher level constituents per word	0.68	0.84	0.76	0.04	0.61	0.80	0.72	0.04	0.66	0.81	0.72	0.03
43 Mean number of words before the main verb of main clause in sentences	1.89	6.63	4.07	1.12	2.09	7.41	4.40	1.30	2.07	11.57	5.23	2.15
44 Type-token ratio for all content words	0.64	0.83	0.75	0.04	0.59	0.95	0.75	0.08	0.59	0.95	0.80	0.07
45 Celex, raw, mean for content words (0–1,000,000)	1,653.73	4,126.18	2,726.08	521.60	1,112.45	4,212.87	2,477.21	704.70	1,007.65	4,741.96	2,518.71	709.13
46 Celex, logarithm, mean for content words (0–6)	2.15	2.65	2.35	0.11	1.94	2.48	2.21	0.15	1.87	2.47	2.18	0.13
47 Celex, raw, minimum in sentence for content words (0–1,000,000)	14.69	322.96	60.61	62.29	3.40	147.33	31.10	30.59	1.33	225.00	36.72	46.08
48 Celex, logarithm, minimum in sentence for content words (0–6)	0.95	1.59	1.22	0.16	0.58	1.50	1.02	0.24	0.30	1.57	0.98	0.26
49 Concreteness, mean for content words	339.48	419.74	380.35	18.81	328.78	426.96	380.71	23.10	327.68	436.06	372.42	26.94
50 Incidence of positive logical connectives	9.42	45.59	23.58	7.34	3.75	37.79	20.43	7.79	3.30	52.15	20.68	9.52
51 Incidence of negative logical connectives	2.93	24.39	12.55	5.15	4.17	30.12	13.49	5.29	0.00	46.69	14.40	7.78
52 Ratio of intentional particles to intentional content	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
53 Incidence of intentional actions, events and particles	7.53	29.81	16.30	5.80	0.00	35.86	12.53	6.68	0.00	32.40	11.21	8.09
54 Mean of tense and aspect repetition scores	0.50	0.97	0.80	0.09	0.60	1.00	0.81	0.09	0.41	0.95	0.79	0.12
55 Sentence syntax similarity, adjacent	0.05	0.13	0.09	0.02	0.04	0.14	0.08	0.02	0.04	0.14	0.08	0.02
56 Sentence syntax similarity, all, across paragraphs	0.05	0.11	0.09	0.02	0.05	0.12	0.08	0.02	0.04	0.12	0.08	0.02
57 Sentence syntax similarity, sentence all, within paragraphs	0.05	0.14	0.09	0.02	0.04	0.14	0.09	0.02	0.03	0.16	0.08	0.02
58 Proportion of content words that overlap between adjacent sentences	0.03	0.15	0.08	0.03	0.01	0.14	0.07	0.03	0.01	0.16	0.07	0.03
59 Mean of location and motion ratio scores	0.31	0.67	0.49	0.07	0.37	0.63	0.51	0.06	0.30	0.74	0.52	0.09
60 Concreteness, minimum in sentence for content words	158.00	194.00	170.92	15.53	158.00	223.00	167.55	16.72	158.00	225.00	181.72	21.35
VocabProfile												
Academic Word List (AWL)	0.00	6.30	1.61	1.26	0.00	5.25	1.64	1.41	0.93	12.96	5.02	2.84
Offlist words	0.00	2.48	0.67	0.59	0.00	3.89	1.05	0.91	0.00	7.68	1.61	1.64
Average characters	3.78	4.65	4.28	0.22	3.69	5.33	4.64	0.34	3.53	5.65	4.86	0.36
British National Corpus (BNC) 1000	76.56	91.69	84.51	3.89	64.93	91.80	78.02	5.77	69.65	89.71	79.70	4.27
BNC 2000	4.10	12.90	7.46	2.32	3.40	13.90	8.80	2.15	3.75	12.67	8.28	1.95
BNC below 2000	85.90	96.97	91.96	2.65	73.43	95.20	86.82	4.79	80.00	95.45	87.98	3.40

indices were considered essential to text selection from our theoretical perspective.

Procedure 2: Comparisons across test levels

The next step was to determine whether observed differences in the values for each index across the three levels were significant ($p < .05$). This would provide further evidence supporting the selection of potentially informative features and indices. An Independent-Samples Kruskal-Wallis Test – a non-parametric test – was used as an alternative to ANOVA since the assumption of equal variance could not be met in our data. The results are displayed in Table 2 (for *Coh-Metrix*) and Table 3 (for *VocabProfile*).

Procedure 3: Posthoc box-plot analyses

To aid interpretation and to provide the focus group (see Procedure 4) with a graphic representation of the relationship between each of the characteristics under consideration and the target level of the material, box-plots were generated displaying the range of values for each index at the three levels (see the Appendix). The bottoms and tops of the boxes are always the 25th and 75th percentile (the lower and upper quartiles, respectively), and the band near the middle of the box is always the 50th percentile (the median). The bottom and top of the whisker extending from each box show respectively the minimum and maximum of the data. Any extreme values which are between one and a half and three box lengths from either end of the box are entered as outliers and are indicated by a small circle.

Procedure 4: Selection of indices

A series of focus group meetings were held to examine the list of prospective indices. The materials used by the group included the definition of each index from *Coh-Metrix* 2.0 (http://*Coh-Metrix*.memphis.edu/*Coh-Metrix*Web2/HelpFile2.htm), the Kruskal-Wallis analyses displayed in Table 2 and Table 3, and the box-plots representing the range of values for each index at the three different levels (see the Appendix).

The focus group set out to identify a core set of indices that might be employed operationally in differentiating between the three levels. They weighed the results of the statistical analysis against the role each index might play in contributing to judgements based on the socio-cognitive validity framework and the Khalifa–Weir (2009) model of reading.

As part of the selection process, the focus group, drawing on our review of the literature, sought to relate each of the selected indices to the cognitive processes described in the socio-cognitive model. Decisions were taken to exclude indices for the following reasons:

1. Lack of any clear relationship to the theoretical framework, e.g. *17 stem overlap adjacent, unweighted*. Indices excluded are: 17, 20, 22, 35, 36.
2. Overlap between indices, e.g., *55 sentence syntax similarity adjacent* showed little difference from *56 sentence syntax similarity, all, across paragraphs*. Indices excluded are: 55, 57.
3. Redundancy where two or more indices involved the same measures, e.g. *40 Flesch-Kincaid Grade Level* and *39 Flesch Reading Ease Score* are both based on measurements of *38 average syllables per word* and *37 average words per sentence*; similarly the *BNC 1000* and *2000* word frequency are based on much the same measurements as the *Celex* indices. Indices excluded are: *CohMx* 38, 40 and *BNC 1000* and *2000*.
4. Issues of interpretation, e.g. some indices are affected by text type. The number of first person narratives employed at *Cambridge English: First* affects the occurrence of *23 ratio of pronouns to noun phrases*. Indices excluded are: *CohMx* 12, 14, 23.
5. Practical concerns, e.g., 28 *LSA sentences all combination mean*. Indices excluded are: *CohMx* 28, 25.
6. Insufficient differentiation between adjacent levels of proficiency. Unsatisfactory or incongruous results on the box-plots led to the rejection of the following indices: 7, 9, 10, 11, 15, 29, 30, 32, 33, 34, 47, 48, 53.

These filtering procedures eliminated a further 30 indices from our study, leaving a set of 17 indices that were felt to be potentially useful in establishing text complexity for operational purposes.

The 17 differentiating indices – Key finding 1

The 17 selected indices were grouped into three categories, namely, lexical, syntactic and text-level representation, thus reflecting different aspects of the cognitive processes identified in the Khalifa and Weir (2009) reading model. The lexical indices are: 38, 42, 44, 46, Academic Word List (AWL) and Offlist words. The syntactic indices are: 27, 37, 41, 43, 56. The text-level representation indices are: 16, 18, 21, 26, 58, 60. Below we provide a description of these indices.

Lexical indices

CohMx38 *Average syllables per word*

The notion that a skilled reader identifies a word purely by its shape has long been discredited. Current models of lexical recognition assume that a reader achieves lexical recognition by drawing upon a number of different cues in parallel (Rastle 2007). A word on the page is matched to an item in the reader's lexicon on the strength of: letter features, letters, digraphs, letter sequences, syllables and the word as a whole. Of these, the units most easily recognised by a computer programme are the syllable and the whole word. Readers take longer to process a multisyllabic word than a monosyllabic one, allowing for frequency effects (Rayner and Pollatsek 1989). The demands of decoding a text at lexical level are thus better measured by counting syllables than by counting whole words.

CohMx42 *Higher level constituents per word*

The mean number of higher level constituents per sentence, controlling for the number of words. The term 'higher-level constituents' is not adequately explained in the *Coh-Metrix* guidance. However, it seems reasonable to assume that it refers to main and subordinate clauses. In fact, there appear to be two issues here: one relating to the number of higher-level constituents per number of words and one to the number of higher-level constituents per sentence.

If reversed, the first of these would indicate the mean length of the clauses in the text, whether main or subordinate. This has implications for the number of words that have to be held

Table 2: Results of Independent-Samples Kruskal-Wallis Test: *Coh-Metrix* measures

	Null hypothesis	Sig.	Decision
7	The distribution of *7 Incidence of causal verbs, links and particles* is the same across categories of levels.	.000	Reject the null hypothesis
8	The distribution of *8 Ratio of causal particles to causal verbs (cp divided by cv+1)* is the same across categories of levels.	.637	**Retain the null hypothesis**
9	The distribution of *9 Incidence of positive additive connectives* is the same across categories of levels.	.000	Reject the null hypothesis
10	The distribution of *10 Incidence of positive temporal connectives* is the same across categories of levels.	.000	Reject the null hypothesis
11	The distribution of *11 Incidence of positive causal connectives* is the same across categories of levels.	.000	Reject the null hypothesis
12	The distribution of *12 Incidence of negative additive connectives* is the same across categories of levels.	.000	Reject the null hypothesis
13	The distribution of *13 Incidence of negative temporal connectives* is the same across categories of levels.	.414	**Retain the null hypothesis**
14	The distribution of *14 Incidence of negative causal connectives* is the same across categories of levels.	.000	Reject the null hypothesis
15	The distribution of *15 Incidence of all connectives* is the same across categories of levels.	.000	Reject the null hypothesis
16	The distribution of *16 Argument overlap, adjacent, unweighted* is the same across categories of levels.	.000	Reject the null hypothesis
17	The distribution of *17 Stem overlap, adjacent, unweighted* is the same across categories of levels.	.000	Reject the null hypothesis
18	The distribution of *18 Anaphor reference, adjacent, unweighted* is the same across categories of levels.	.000	Reject the null hypothesis
19	The distribution of *19 Argument overlap*, all distances is the same across categories of levels.	.559	**Retain the null hypothesis**
20	The distribution of *20 Stem overlap, all distances, unweighted* is the same across categories of levels.	.005	Reject the null hypothesis
21	The distribution of *21 Anaphor reference, all distances* is the same across categories of levels.	.000	Reject the null hypothesis
22	The distribution of *22 Noun phrase incidence score (per thousand words)* is the same across categories of levels.	.000	Reject the null hypothesis
23	The distribution of *23 Ratio of pronouns to noun phrases* is the same across categories of levels.	.000	Reject the null hypothesis
24	The distribution of *24 Number of conditional expressions, incidence score* is the same across categories of levels.	.730	**Retain the null hypothesis**
25	The distribution of *25 Number of negations, incidence score* is the same across categories of levels.	.015	Reject the null hypothesis
26	The distribution of *26 Logical operator incidence score* is the same across categories of levels.	.001	Reject the null hypothesis
27	The distribution of *27 LSA, sentence to sentence adjacent mean* is the same across categories of levels.	.000	Reject the null hypothesis
28	The distribution of *28 LSA, sentences all combinations mean* is the same across categories of levels.	.001	Reject the null hypothesis
29	The distribution of *29 LSA, paragraph to paragraph mean* is the same across categories of levels.	.001	Reject the null hypothesis
30	The distribution of *30 Personal pronoun incidence score* is the same across categories of levels.	.000	Reject the null hypothesis
31	The distribution of *31 Mean hypernym values of nouns* is the same across categories of levels.	.062	**Retain the null hypothesis**
32	The distribution of *32 Mean hypernym values of verbs* is the same across categories of levels.	.024	Reject the null hypothesis
33	The distribution of *33 Number of paragraphs* is the same across categories of levels.	.000	Reject the null hypothesis
34	The distribution of *34 Number of sentences* is the same across categories of levels.	.000	Reject the null hypothesis
35	The distribution of *35 Number of words* is the same across categories of levels.	.000	Reject the null hypothesis
36	The distribution of *36 Average sentences per paragraph* is the same across categories of levels.	.289	**Retain the null hypothesis**
37	The distribution of *37 Average words per sentence* is the same across categories of levels.	.000	Reject the null hypothesis
38	The distribution of *38 Average syllables per word* is the same across categories of levels.	.000	Reject the null hypothesis
39	The distribution of *39 Flesch Reading Ease Score (0–100)* is the same across categories of levels.	.000	Reject the null hypothesis
40	The distribution of *40 Flesch-Kincaid Grade Level (0–12)* is the same across categories of levels.	.000	Reject the null hypothesis
41	The distribution of *41 Mean number of modifiers per noun phrase* is the same across categories of levels.	.000	Reject the null hypothesis
42	The distribution of *42 Higher level constituents per word* is the same across categories of levels.	.000	Reject the null hypothesis
43	The distribution of *43 Mean number of words before the main verb of main clause in sentences* is the same across categories of levels.	.014	Reject the null hypothesis
44	The distribution of *44 Type-token ratio for all content words* is the same across categories of levels.	.000	Reject the null hypothesis
45	The distribution of *45 Celex, raw, mean for content words (0–1,000,000)* is the same across categories of levels.	.052	**Retain the null hypothesis**
46	The distribution of *46 Celex, logarithm, mean for content words (0–6)* is the same across categories of levels.	.000	Reject the null hypothesis
47	The distribution of *47 Celex, raw, minimum in sentence for content words (0–1,000,000)* is the same across categories of levels.	.000	Reject the null hypothesis
48	The distribution of *48 Celex, logarithm, minimum in sentence content words (0–6)* is the same across categories of levels.	.000	Reject the null hypothesis
49	The distribution of *49 Concreteness, mean for content words* is the same across categories of levels.	.082	**Retain the null hypothesis**
50	The distribution of *50 Incidence of positive logical connectives* is the same across categories of levels.	.085	**Retain the null hypothesis**
51	The distribution of *51 Incidence of negative logical connectives* is the same across categories of levels.	.506	**Retain the null hypothesis**
52	The distribution of *52 Ratio of intentional particles to intentional content* is the same across categories of levels.	1.000	**Retain the null hypothesis**

Table 2: (continued)

	Null hypothesis	Sig.	Decision
53	The distribution of *53 Incidence of intentional actions, events and particles* is the same across categories of levels.	.000	Reject the null hypothesis
54	The distribution of *54 Mean of tense and aspect repetition scores* is the same across categories of levels.	.778	**Retain the null hypothesis**
55	The distribution of *55 Sentence syntax similarity, adjacent* is the same across categories of levels.	.039	Reject the null hypothesis
56	The distribution of *56 Sentence syntax similarity, all, across paragraphs* is the same across categories of levels.	.003	Reject the null hypothesis
57	The distribution of *57 Sentence syntax similarity, sentence all, within paragraphs* is the same across categories of levels.	.005	Reject the null hypothesis
58	The distribution of *58 Proportion of content words that overlap between adjacent sentences* is the same across categories of levels.	.007	Reject the null hypothesis
59	The distribution of *59 Mean of location and motion ratio scores* is the same across categories of levels.	.170	**Retain the null hypothesis**
60	The distribution of *60 Concreteness, minimum in sentence for content words* is the same across categories of levels.	.000	Reject the null hypothesis

Table 3: Results of Independent-Samples Kruskal-Wallis Test: *VocabProfile* measures

Null hypothesis	Sig.	Decision
The distribution of *AWL* is the same across categories of Levels.	.000	Reject the null hypothesis
The distribution of *Offlist words* is the same across categories of Levels.	.001	Reject the null hypothesis
The distribution of *average characters* is the same across categories of Levels.	.000	Reject the null hypothesis
The distribution of *BNC 1000* is the same across categories of Levels.	.000	Reject the null hypothesis
The distribution of *BNC 2000* is the same across categories of Levels.	.006	Reject the null hypothesis
The distribution of *BNC below 2000* is the same across categories of Levels.	.000	Reject the null hypothesis

in mind during parsing since it is usually at clause boundaries that strings of words are 'made up' into propositions (Jarvella 1971). The second provides an indication of the extent to which a text contains embedded clauses. Main verbs in a sentence are broadly indicative of the number of clauses. Sentences with complex syntactic composition have one or more clauses embedded in them and therefore have a higher incidence of verb phrases. Clearly, the higher the ratio of clauses to sentences, the higher the likelihood that a sentence will contain subordinate clauses. Subordinate clauses increase processing demands because, within the domain of a single sentence, a reader has to parse multiple groups of words into propositions and then to trace conceptual and logical links between the propositions that have been derived. This is clearly much more demanding than processing a series of simple sentences.

CohMx44 *Type-token ratio for all content words*

Type-token ratio (TTR) (Templin 1957) is the number of unique words (called types) divided by the number of tokens of these words. Each unique word in a text is considered a word type. Each instance of a particular word is a token. For example, if the word 'dog' appears in the text seven times, its type value is 1, whereas its token value is 7. When the type-token ratio approaches 1, each word occurs only once in the text; comprehension should be comparatively difficult because many unique words need to be decoded and integrated with the discourse context. As the type-token ratio decreases, words are repeated many times in the text, which should increase the ease and speed of text processing. Type-token ratios are computed for content words, but not function words.

As the length of the reading passage increases, so does the probability that any given word will recur. However, the range of word types also increases. In written English, stylistic constraints militate against the repetition of words in adjacent sentences. To avoid duplication, writers will often exercise a preference for a synonym rather than a pro-form – thus increasing the TTR. When tests target higher level language learners, authentic textual considerations of this kind come increasingly into play. Because type-token ratios are sensitive to text length, where text length varies substantially, standardised values should be employed. As the texts in this study vary in length, a standardised type-token ratio was employed.

CohMx46 *Celex, logarithm, mean for content words (0–6)*

This is the logarithm of the frequency of all content words in the text. Taking the log of the frequencies rather than the raw scores is compatible with research on reading time (Just and Carpenter 1980), on automatic decoding as a strong predictor of L2 reading performance (Koda 2005) and so this *Celex* measure obtained through *Coh-Metrix* was preferred in principle to the percentages provided for different frequency levels obtained through *VocabProfile*. The word with the lowest log frequency score is the least frequent word in the sentence. Texts containing a high proportion of low-frequency words will be more difficult to process than those containing only very common words.

Less frequent words are associated with slower decoding times (Garman 1985). In addition, a high ratio of low-frequency content words increases the likelihood that a passage will contain a number of words that are unfamiliar to the test taker. However, not too much should be made of

the contribution made by unfamiliar words to text difficulty. The fact is that many such words can be decoded by using analogy or derivational morphology; others can be ignored as not central to the main argument of the text. The issue determining difficulty is not necessarily low frequency, but rather the transparency of the words.

AWL (Academic Word List)

This is the percentage of words in a text also appearing on the AWL (sub-technical vocabulary). The academic word lists identify words used more commonly in academic than in other contexts, particularly the sub-technical vocabulary that occurs across disciplines (Campion and Elley 1971, Coxhead 2000).

The difference between general and academic vocabulary would seem to involve a higher level of abstractness, which overlaps with index *CohMx60 Concreteness, minimum in sentence for content words*. In addition, there would seem to be considerable overlap between the use of specialised vocabulary and the low-frequency vocabulary specified in index *CohMx46 Celex, logarithm, mean for content words (0–6)*.

Offlist words

These are words that fall outside the most frequent 15,000 words of the British National Corpus and that do not occur on the AWL. These are most likely to be technical words or proper nouns. Such words are unlikely to be familiar to L2 speakers and so may be a source of difficulty. However, as an index of text difficulty, Offlist words, like the AWL, may be redundant because it correlates with other indices such as *CohMx46 Celex, logarithm, mean for content words (0–6)* and *CohMx 38 Average syllables per word*.

Syntactic indices

For *CohMx41* and *CohMx43*, the reader is referred to the description provided in the literature review in this article.

CohMx27 *LSA, sentence to sentence adjacent mean*

Mean Latent Semantic Analysis (LSA) cosines for adjacent, sentence to sentence measure how conceptually similar each sentence is to the next sentence. Text cohesion, which facilitates reading, is assumed to increase as a function of higher cosine scores between text constituents.

CohMx37 *Average words per sentence*

This index appears to be a rough measure of both the syntactic complexity and the lexical density of a sentence. Clearly, the number of words in a sentence must often correlate loosely with the sentence's complexity in terms of number of clauses, and a longer sentence might be denser in lexical terms. This measure partly relates to processing at the level of structure building (Gernsbacher 1990) in that the more complex the sentence, the more elaborate is the structure that has to be assembled. If one assumes that longer sentences might also result from longer and more densely packed clauses, then the measure is also an indicator of difficulty of parsing. In parsing, a reader has to hold a series of words in the mind until such time as they reach the end of a clause and can trace a syntactic pattern in the string (Rayner and Pollatsek 1989). The longer the clause, the more words the reader has to hold in the mind. Lewis, Vasishth and Van Dyke (2006) suggest that processing items towards the end of longer sentences will be harder, since they usually have to be integrated with items that have occurred earlier on in the sentence. Graesser et al (2011) also suggest that longer sentences tend to place more demands on working memory and are therefore more difficult to process.

Khalifa and Weir (2009) describe how sentence length in Cambridge English Reading examinations increases according to the level of the examination, although again there seems to be considerable variation in the lengths of sentences featuring in the tests even at the same level. Again attention to this index might ensure greater homogeneity between the texts used at a particular level.

CohMx56 *Sentence syntax similarity, all, across paragraphs*

The *Sentence syntax similarity* indices in *Coh-Metrix* compare the syntactic tree structures of sentences. An issue is what is known as a syntactic priming effect. It is well attested in language production research (Pickering and Branigan 1999) that after a speaker has formulated a particular syntactic structure, there is a likelihood that they will employ a similar structure in the following utterance. The phenomenon is less clearly attested in reading comprehension. While syntactic priming appears to play a positive role in comprehension, it has been suggested that the effect may be partly or wholly due to the repetition of the verb. However, recent neurological evidence (Ledoux, Traxler and Saab 2007) suggests that syntactic parsing effects may be present even when the verb is not repeated.

Text-level representation indices

CohMx16 *Argument overlap, adjacent, unweighted*

This index is the proportion of all sentence pairs per paragraph that share one or more arguments (i.e. noun, pronoun, noun-phrase). A higher score is indicative of a more cohesive text and easier reading (see the literature review).

CohMx18 *Anaphor reference, adjacent, unweighted*

This is the proportion of anaphor references between adjacent sentences. It is easier to resolve anaphoric reference where the anaphor occurs in a sentence that follows immediately after the one in which the referent occurs. The referent will remain foregrounded in the mind of the reader, and may have been tagged as a current topic focus.

A potential weakness of a simple measure of proximity is that there are occasional cases of ambiguity where the first sentence contains more than one possible referent. In this case, the preferential choice would be made on the basis of parallel function (Arnold, Eisenband, Brown-Schmidt and Trueswell 2000, Sheldon 1974) with the anaphor matched by similarity of sentence function and position (a subject pronoun to a preceding NP subject, an object pronoun to a preceding NP object), rather than by closest proximity to the referent. It is not made explicit whether *Coh-Metrix* excludes from consideration the first and second personal pronouns, which have referents outside the text. Nor is it made explicit whether it includes referents such as *this, that, the former, the latter*.

CohMx21 *Anaphor reference, all distances*

This index measures the proportion of unweighted anaphor references that refer back to a constituent up to five sentences earlier. This would presumably include instances already counted under adjacent sentences (index 18). A more informative measure (with only an incremental effect on difficulty) might be of cases where an anaphoric referent occurs earlier than in the immediately preceding sentence. Where the referent is 'remote' from the anaphor in this way, it is more difficult to process. Indeed, children who are inexperienced readers have problems in resolving this type of anaphor, but are often capable of resolving anaphoric reference where the referent is adjacent (Yuill and Oakhill 1991). The difficulty lies in the need to carry forward one or more current topics, while at the same time, decoding and parsing written text.

CohMx26 *Logical operator incidence score*

This index is the incidence of logical operators, including *and, or, not, if, then* and a small number of other similar cognate terms. Texts with a high density of these logical operators or connectives tend to be more difficult to access. In fact, many of the connectors listed in the guidance for *Coh-Metrix* users do not appear to be logical in function. It is curious that the examples include 'negations' and 'counterfactuals', which are known to be semantically difficult to process in their own right.

Where there is no connective linking adjacent clauses or sentences, the reader has to rely upon inference (Brown and Yule 1983, Oakhill and Garnham 1988, Singer 1994) in order to trace a connection. If there is a logical connector, it marks the relationship between the two idea units unambiguously, and spares the reader the cognitive effort associated with having to infer the connection. This would suggest that the presence of connectives reduces difficulty rather than increasing it (as *Coh-Metrix* seems to suggest). An explanation for the assumption that this measure correlates with difficulty may be that the word incidence in the specification refers to types not to tokens. One would certainly expect a greater range of logical connectives in more advanced texts.

CohMx58 *Proportion of content words that overlap between adjacent sentences*

Crossley, Greenfield et al (2008:483) explain that 'overlapping vocabulary has been found to be an important aspect in reading processing and can lead to gains in text comprehension and reading speed'. The occurrence of the same content word in adjacent sentences reduces text difficulty in two ways. At the level of decoding, a word is subject to a repetition priming effect (Scarborough, Cortese and Scarborough 1977, Stanners, Neiser, Hernin and Hall 1979), whereby a) it is recognised more readily on its second occurrence and b) lexical access is speeded up. At a discourse level, the repetition contributes to text cohesion, thus reinforcing current themes. Repetition priming is surprisingly long-lived and assists a reader in processing recurrent words throughout a text. On the other hand there are stylistic constraints which operate against the use of identical words in adjacent sentences (see *CohMx44 Type-token ratio for all content words*), and foster the use of pro-forms and synonyms.

CohMx60 *Concreteness, minimum in sentence for content words*

For each sentence in the text, a content word is identified that has the lowest concreteness rating. This score is the mean of these low-concreteness words across sentences.

The concern here is with the extent to which the information in a text concerns observable, concrete phenomena (concrete content), unobservable phenomena such as social institutions (abstract content) or, at a higher level of abstraction, theoretical treatments of abstract phenomena (meta-phenomenal content) (Moore and Morton 1999). Different levels of abstraction may, of course, be found within a single text. Alderson et al (2004:127) mark abstractness as a useful feature to consider in estimating text difficulty in relation to the CEFR. Information that is more abstract may prove to be more difficult to process and so divert cognitive resources from language processing. At the same time abstract information often implies a linguistic complexity that may further stretch the L2 reader's resources. Much academic text, particularly in the humanities and social sciences, is concerned with abstract ideas.

This *Coh-Metrix* measure is based upon the well-established finding that abstract words are more difficult to process because they are not as imageable as concrete words. There is some evidence (Bleasdale 1987) that there may be separate lexicons for the two types. However, the measure used draws upon the MRC Psycholinguistic Database which was quite small, compiled a long time ago (Coltheart 1981) and limited in coverage. The abstractness ratings of the words were partly based on a study by Paivio, Yuille and Madigan (1968) which featured only 925 items, but was later expanded to 4,000. Furthermore, the dataset does not deal adequately with a major area of controversy in relation to abstractness – the difference (Kintsch 1972) between abstract words which have been derived morphologically (*happiness*) and others which are unitary (*truth*). A word such as *friendship* would qualify semantically as abstract but can easily be deconstructed through knowledge of derivational suffixation and is closely linked morphologically to its concrete stem.

The 17 differentiating indices – Key finding 2

Multiple regression analysis was used to indicate which of the indices are good predictors of text level. Each set of the indices were entered as independent variables with the text level (*Cambridge English: First*, *Cambridge English: Advanced* and *Cambridge English: Proficiency*) as the dependent variable.

Lexical set as predictor of text level

The lexical indices provided a moderate level of prediction ($R = .669$, adjusted $R^2 = .434$, *F change* (df 1,161) = 4.235, $p = .000$). In terms of individual relationships, four indices emerged as significant ($p < .05$) predictors of text level. These were *AWL* ($t = 7.474$, $p = .000$); *Offlist* ($t = 3.990$, $p = .000$); *CohMx46 Celex, logarithm, mean for content words (0–6)* ($t = -3.990$, $p = .003$) and *CohMx44 Type-token ratio for all content words* ($t = 2.058$, $p = .041$).

Syntactic set as predictor of text level

The syntactic indices were considerably less effective than the lexical indices as predictors of text level ($R = .394$, adjusted $R^2 = .145$, F change (df 1,163) = 6.930, $p = .009$). At the index level, two of the five indices entered to the model emerged as significant predictors: *CohMx37 Average words per sentence* ($t = .255$, $p = .002$) and *CohMx41 Mean number of modifiers per noun phrase* ($t = .210$, $p = .009$).

Text level representation set as predictor of text level

Although less effective than the lexical indices, the cohesion and concreteness indices also proved moderately effective as predictors of text level ($R = .590$, adjusted $R^2 = .328$, F change (df 1,160) = 14.996, $p = .000$). Individual indices that emerged as predictive of text level included: *CohMx21 Anaphor reference, all distances* ($t = .255$, $p = .002$); *CohMx26 Logical operator incidence score* ($t = 3.758$, $p = .000$); *CohMx60 Concreteness, minimum in sentence for content words* ($t = 3.033$, $p = .003$); *CohMx16 Argument overlap, adjacent, unweighted* ($t = -4.190$, $p = .000$); and *CohMx58 Proportion of content words that overlap between adjacent sentences* ($t = 3.872$, $p = .000$).

From these analyses it is clear that, although each set had some predictive power, no single set of indices provided a very strong indication of whether a text was associated with *Cambridge English: First*, *Cambridge English: Advanced* or *Cambridge English: Proficiency*. Next, indices which had been significant ($p < .05$) predictors of text level when the three feature sets were treated separately were combined into a single predictor set and used together as the independent variables in a further multiple regression analysis. The level of prediction (adjusted $r^2 = .583$) improved substantially on that achieved by any one set of indices ($R = .782$, adjusted $R^2 = .537$, F change (df 11, 154) = 21.980, $p = .000$). At the individual level, the following indices emerged as being predictive of text level: *CohMx26 Logical operator incidence score* ($t = 3.076$, $p = .002$); *CohMx16 Argument overlap, adjacent, unweighted* ($t = -5.946$, $p = .000$); *CohMx58 Proportion of content words that overlap between adjacent sentences* ($t = 4.116$, $p = .000$); *CohMx46 Celex, logarithm, mean for content words (0–6)* ($t = 3.043$, $p = .003$); *AWL* ($t = 5.377$, $p = .000$); and *Offlist* ($t = 3.224$, $p = .002$).

Summary

Multiple regression analysis suggests that features of cohesion (logical operator incidence, lexical overlap between sentences) and lexis (word frequency and the occurrence of infrequent and academic words) rather than syntax are criterial in distinguishing between the texts used in the three highest levels of the Cambridge English examinations.

Conclusion

The automated approach to text analysis for language testing purposes is obviously still in its infancy and it will take some time to be able to confirm whether the indices we selected in this initial study are the most appropriate. Nevertheless, the indices chosen represent a principled attempt to establish a set of parameters for operational text selection. Adhering to these parameters should improve the consistency of test material over time.

It is hoped that this study has generated valuable insights to set alongside previous findings from *Coh-Metrix* studies, providing useful additional justification for the validity of the indices and for the application of computational tools, such as *Coh-Metrix*, to analyse text characteristics.

References

Alderson, J C (2000) *Assessing Reading*, Cambridge: Cambridge University Press.

Alderson, J C, Figueras, N, Kuijper, H, Nold, G, Takala, S and Tardieu, C (2004) *The development of specifications for item development and classification within the common European framework of reference for languages: learning, teaching, assessment. Reading and listening, The final report of the Dutch CEF Construct Project,* Project Report, Lancaster University, Lancaster, available online: eprints.lancs.ac.uk/44

Arnold, J E, Eisenband, J, Brown-Schmidt S, and Trueswell, J C (2000) The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking, *Cognition* 76, 13–26.

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

Bachman, L F, Davidson, F, Ryan, K and Choi, I (1995) *An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL Comparability Study*, Studies in Language Testing volume 1, Cambridge: UCLES/Cambridge University Press.

Barker, K (1998) A trainable bracketer for noun modifiers, in *Proceedings of the Twelfth Canadian Conference on Artificial Intelligence (LNAI 1418)*, KAML Group: Vancouver, 196–210.

Bleasdale, F A (1987) Concreteness dependent associative priming: Separate lexical organization for concrete and abstract words, *Journal of Experimental Psychology: Learning, Memory, & Cognition* 13, 582–594.

Brown, G T L and Yule, G (1983) *Discourse Analysis*, Cambridge: Cambridge University Press.

Brown, J D (1997) An EFL readability index, *University of Hawaii Working Papers in English as a Second Language* 15 (2), 85–119.

Campion, M E and Elley, W B (1971) *An Academic Vocabulary List*, Wellington: NZCER.

Cobb, T (2002) *Web Vocabprofile,* available online: www.lextutor.ca/vp/

Cobb, T (2003) *VocabProfile, The Compleat Lexical Tutor*, available online: www.lextutor.ca

Coltheart, M (1981) The MRC psycholinguistic database, *Quarterly Journal of Experimental Psychology* 33A, 497–505.

Coxhead, A (2000) A new academic word list, *TESOL Quarterly* 34 (2), 213–238.

Crossley, S A, Greenfield, J and McNamara, D S (2008) Assessing text readability using cognitively based indices, *TESOL Quarterly* 42 (3), 475–493.

Crossley, S A, Louwerse, M M and McNamara, D S (2008) Identifying linguistic cues that distinguish text types: A comparison of first and second language speakers, *Language Research* 44 (2), 361–381.

Enright, M, Grabe, W, Koda, K, Mosenthal, P, Mulcahy-Ernt, P and Schedl, M (2000) *TOEFL 2000 Reading Framework: A Working Paper*, TOEFL Monograph Series 17, Princeton: Educational Testing Service.

Fortus, R, Coriat, R and Fund, S (1998) Prediction of item difficulty in the English subtest of Israel's inter-university psychometric entrance test, in Kunnan, A J (Ed.) *Validation in Language Assessment: Selected Papers from the 17th Language Research Colloquium, Long Beach*, Mahwah: Lawrence Erlbaum, 61–87.

Freedle, R and Kostin, I (1993) *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item*

types: Main idea, inference, and supporting idea items, TOEFL Research Reports No. RR-93-44, Princeton: Educational Testing Service.

Garman, A (1985) *Psycholinguistics: Central Topics*, London: Methuen.

Gernsbacher, M A (1990) *Language Comprehension as Structure Building*, Mahwah: Erlbaum.

Gervasi, V and Ambriola, V (2002) Quantitative assessment of textual complexity, in Merlini Barbesi, L (Ed.) *Complexity in Language and Text*, Pisa: PLUS-University of Pisa, 197–228.

Goldman, S and Rakestraw, J (2000) Structural aspects of constructing meaning from text, in Kamil, M, Rosenthal, P, Pearson, P and Barr, R (Eds) *Handbook of Reading Research*, Mahwah: Lawrence Erlbaum, 331–335.

Graesser, A C, McNamara, D S, Louwerse, M M and Cai, Z (2004) Coh-Metrix: Analysis of text on cohesion and language, *Behavioral Research Methods, Instruments, and Computers* 36, 193–202.

Graesser, A C, McNamara, D S and Kulikowich, J (2011) Coh-Metrix: Providing multilevel analyses of text characteristics, *Educational Researcher* 40 (5), 223–234.

Green, A (2012) *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range*, English Profile Studies volume 2, Cambridge: UCLES/Cambridge University Press.

Green, A, Ünaldi, A and Weir, C J (2010) Empiricism versus connoisseurship: establishing the appropriacy of texts for testing reading for academic purposes, *Language Testing* 27 (3), 1–21.

Jarvella, R J (1971) Syntactic processing of connected speech, *Journal of Verbal Learning and Verbal Behavior* 10, 409–416.

Just, M A and Carpenter, P A (1980) A theory of reading: From eye fixations to comprehension, *Psychological Review* 87 (4), 329–354.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

Kintsch, W (1972) Abstract nouns: Imagery versus lexical complexity, *Journal of Verbal Learning and Verbal Behaviour* 11, 59–65.

Koda, K (2005) *Insights into Second Language Reading: A Cross-linguistic Approach*, Cambridge: Cambridge University Press.

Ledoux, K, Traxler, M J and Saab, T Y (2007) Syntactic priming in comprehension: Evidence from event-related potentials, *Psychological Science* 18, 135–143.

Lewis, L R, Vasishth, S and Van Dyke, J (2006) Computational principles of working memory in sentence comprehension, *Trends in Cognitive Science* 10, 447–454.

Masi, S (2002) The literature on complexity, in Merlini Barbesi, L (Ed.) *Complexity in Language and Text*, Pisa: PLUS-University of Pisa: 197–228.

McNamara, D S, Louwerse, M M, Cai, Z and Graesser, A C (2005) *Coh-Metrix version 1.4*, available online: Coh-Metrix.memphis.edu.

McNamara, D S, Graesser, A C and Louwerse, M M (in press) Sources of text difficulty: Across the ages and genres, in Sabatini, J P and Albro, E (Eds) *Assessing Reading in the 21st Century: Aligning and Applying Advances in the Reading and Measurement Sciences*, Lanham: R and L Education.

Moore, T and Morton, J (1999) Authenticity in the IELTS Academic Module Writing test: A comparative study of Task 2 items and university assignments, in Tulloh, R (Ed.) *IELTS Research Reports Volume 2*, Canberra: IELTS Australia, 64–106.

Oakhill, J V and Garnham, A (1988) *Becoming a Skilled Reader*, Oxford: Basil Blackwell.

Ortega, L (2003) Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing, *Applied Linguistics* 24 (4), 492–18.

Paivio, A, Yuille, J C and Madigan, S A (1968) Concreteness, imagery and meaningfulness values for 925 nouns, *Journal of Experimental Psychology Monograph Supplement* 76 (1), 1–25.

Perfetti, C A, Landi, N and Oakhill, J (2005) The acquisition of reading comprehension skill, in Snowling, M J and Hulme, C (Eds) *The Science of Reading: A Handbook*, Oxford: Blackwell, 227–257.

Pickering, M J and Branigan, H P (1999) Syntactic priming in language production, *Trends in Cognitive Sciences* 3 (4), 136–141.

Rastle, K (2007) Visual word recognition, in Gaskell, M G (Ed.) *The Oxford Handbook of Psycholinguistics*, Oxford: Oxford University Press, 71–87.

Rayner, K and Pollatsek, A (1989) *The Psychology of Reading*, Eaglewood Cliffs: Prentice Hall.

Scarborough, D L, Cortese, C and Scarborough, H S (1977) Frequency and repetition effects in lexical memory, *Journal of Experimental Psychology: human perception and performance* 3, 1–17.

Sheldon, A (1974) The role of parallel function in the acquisition of relative clauses in English, *Journal of Verbal Learning and Verbal Behaviour* 13, 272–281.

Singer, M (1994) Discourse inference processes, in Gernsbacher, M A (Ed.) *Handbook of Psycholinguistics*, San Diego: Academic Press, 479–516.

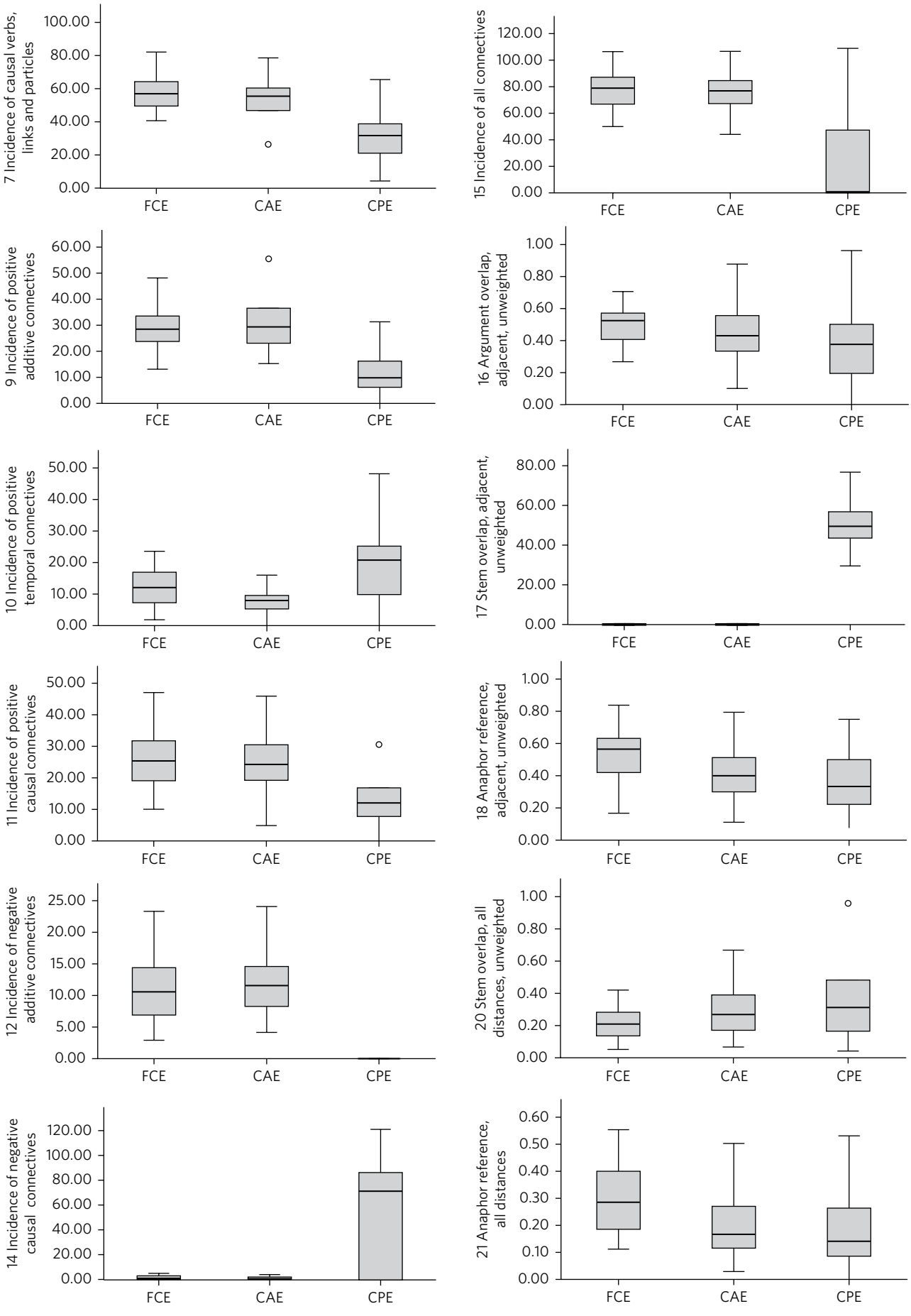
Stanners, R F, Neiser, J J, Hernin, W P and Hall, R (1979) Memory representation for morphologically related words, *Journal of Verbal Learning and Verbal Behaviour* 18, 399–412.

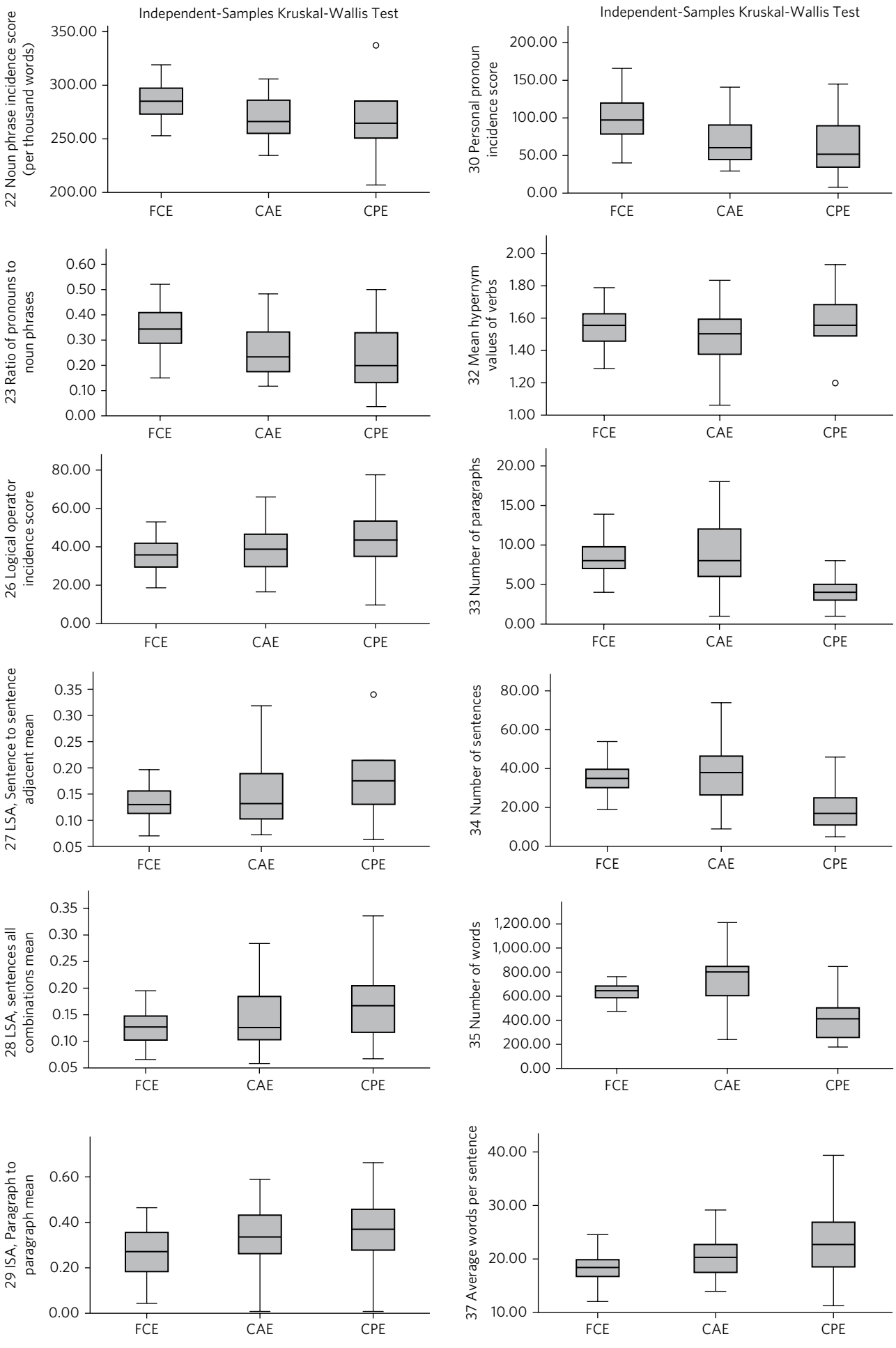
Templin, M (1957) *Certain Language Skills in Children*, Minneapolis: University of Minnesota Press.

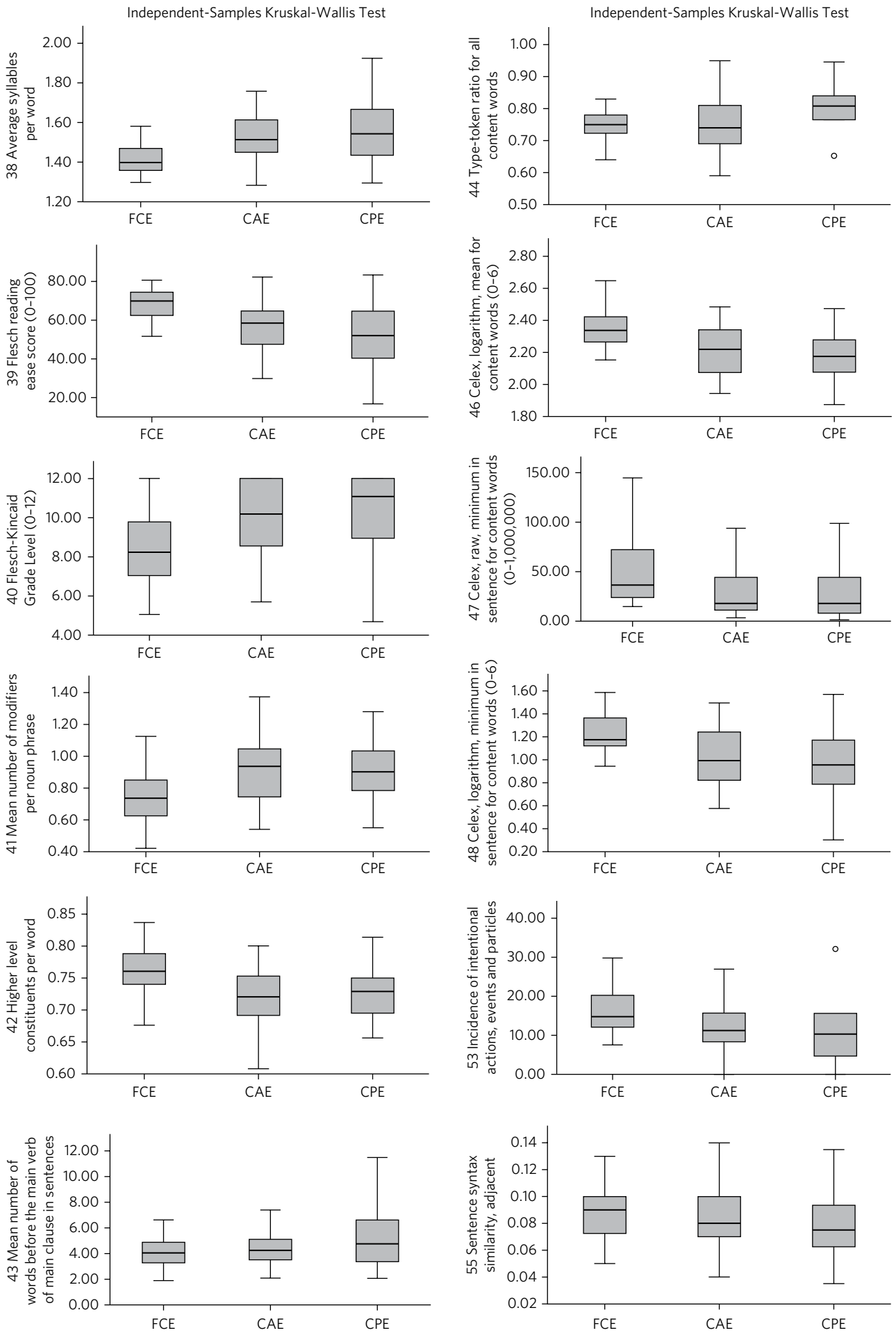
Wolfe-Quintero, K, Inagaki, S and Kim, H-Y (1998) *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*, Honolulu: University of Hawaii Press.

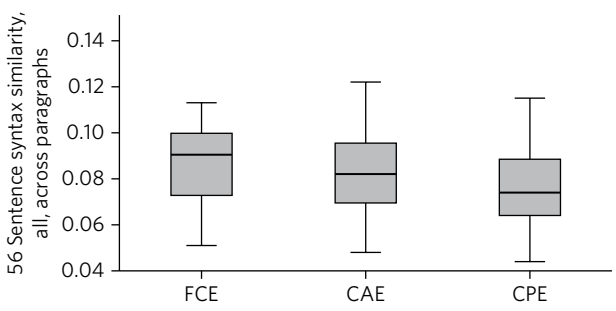
Yuill, N and Oakhill, J (1991) *Children's Problems in Text Comprehension*, Cambridge: Cambridge University Press.

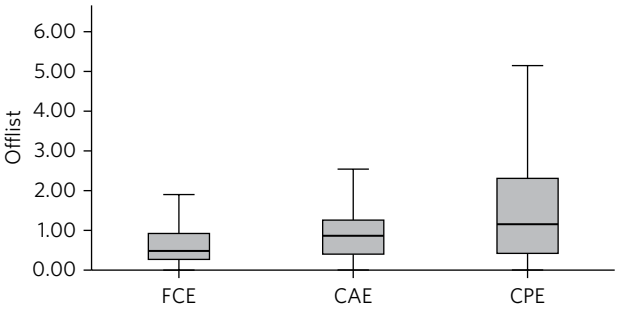
Appendix: Box plots of text analysis indices at three levels

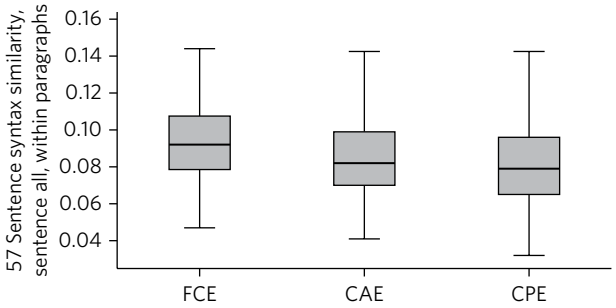


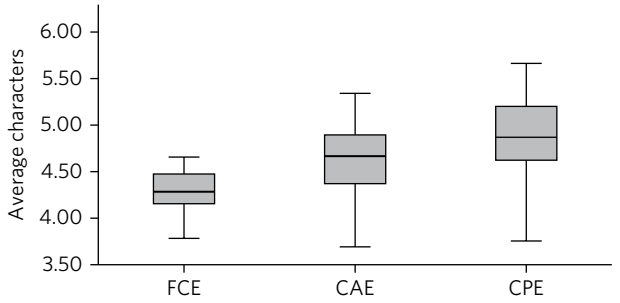


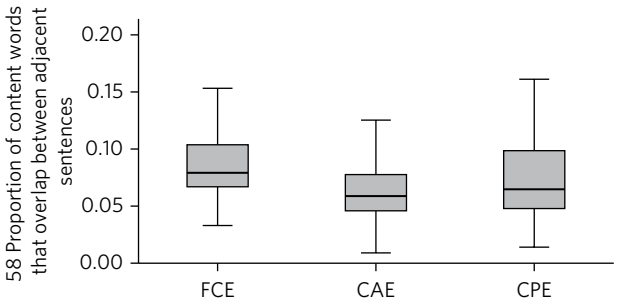


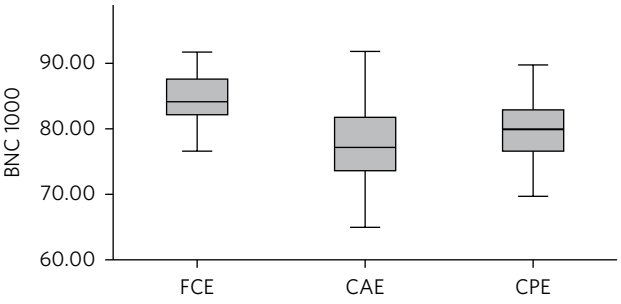


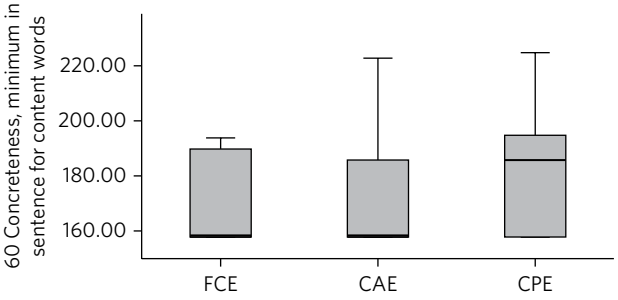


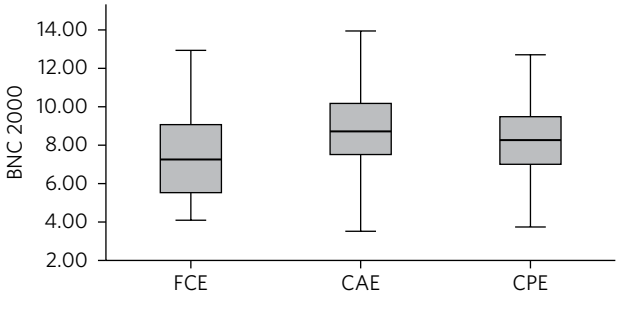


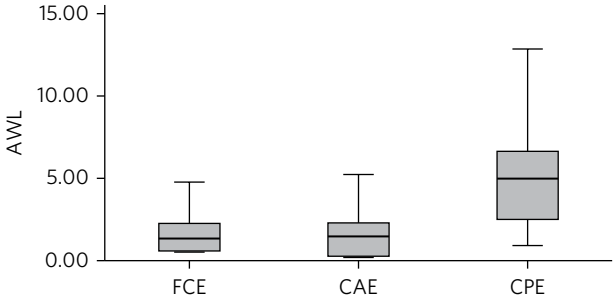


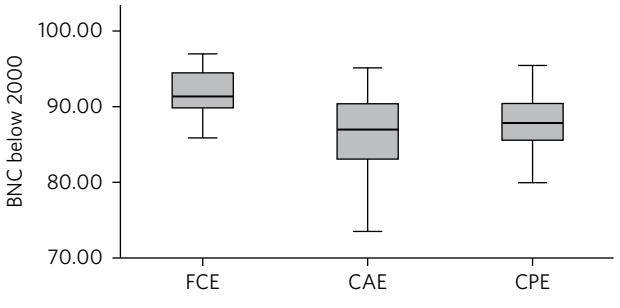












Linguistic analysis of speaking features distinguishing general English exams at CEFR levels

OKIM KANG ENGLISH DEPARTMENT, NORTHERN ARIZONA UNIVERSITY, USA

Introduction

Cambridge English Speaking tests reflect a view of speaking ability that involves multiple competences such as lexico-grammatical knowledge, phonological control, and pragmatic awareness (Taylor 2001, 2011). The competences represent descriptive features of language use (Davies 2008) and take account of cognitive descriptions of the speech production process (Field 2011, Garman 1990). However, relatively few studies have addressed the linguistic characteristics of examinees' production from the perspective of second language (L2) speaking assessment. For example, research such as Iwashita, Brown, McNamara and O'Hagan (2008) attempted to demonstrate speaking features associated with oral assessment by using a range of measures such as grammatical accuracy and complexity, vocabulary, pronunciation and fluency. They found that a certain set of features had an impact on the overall/holistically assigned score with particular features of vocabulary and fluency having the strongest effect. However, their findings were still unclear with regard to specific linguistic characteristics that differentiated L2 speakers at various proficiency levels. Therefore, the conjoined impact of a wide array of linguistic features on Cambridge English proficiency judgements of candidate speech needs to be further investigated as a validation process of L2 oral proficiency tests.

This study involved Cambridge English examinations ranging from B1 to C2, namely, *Cambridge English: Preliminary* (*PET*), *Cambridge English: First* (*FCE*), *Cambridge English: Advanced* (*CAE*), and *Cambridge English: Proficiency* (*CPE*). The *Cambridge English: Key* (*KET*) exam (A2) has been excluded in this project because of the interest of the research in the component of an examinee's long uninterrupted turn, which is a task type not present in this exam. The specific aims of this study are to examine what speaking features can distinguish examinees' performances at different proficiency levels in general English exams based on assessment criteria (grammatical and lexical resources, discourse management and pronunciation).

Background

Validity has been considered as a crucial concept in language testing and assessment. Stemming from an argument-based method, the current validity approach moves towards developing interpretive validity arguments (e.g. Kane 1992). That is, it starts from the *grounds* represented by an observation of test takers' performance on a test. Then, a *conclusion* of a test taker's ability is drawn from the observation, based on a chain of reasoning which includes *inferences* and their *backing* (usually manifested by empirical research). This validity framework is made up of a set of inferences (e.g. domain description, evaluation, generalisation, explanation, extrapolation and utilisation) moving the argument from the grounds to the conclusion (Chapelle, Enright and Jamieson 2008). This study focuses on two inferences: evaluation (i.e. evaluating observed scores that reflect targeted language abilities) and explanation (i.e. warranting that expected scores are attributed to a construct of general language proficiency).

Increasingly over the last decade, candidate speaking performance has undergone linguistic analysis in order to investigate distinguishing features at different proficiency levels (e.g. TOEFL or *IELTS*). Linguistic features often include aspects of vocabulary, grammar, fluency, content, pronunciation and rhetorical organisation (e.g. Brown, Iwashita and McNamara 2005, Iwashita et al 2008). In Cambridge English exams evaluation criteria, L2 speaking proficiency is represented by the following dimensions: discourse management, lexical resource, grammatical resource and pronunciation. Detailed linguistic analyses are discussed in the methodology section. Note that the current report excludes the dimension of *interactive communication*, which is also part of the Cambridge English evaluation criteria, because the speech samples used in this project solely focused on the mono-logic component of the test.

Research question

The project is guided by the following research question:

- What are the overall salient linguistic features that distinguish speaking Levels B1–C2 on the Common European Framework of Reference for Languages (CEFR) in Cambridge English examinations for the following scoring criteria: (a) discourse management (b) grammatical and lexical resources and (c) pronunciation?

Methodology

Speech files

Cambridge English Language Assessment provided 120 audio speech files of examinee responses from general English examinations at CEFR Levels B1 to C2 as follows: *Cambridge English: Preliminary* (32), *Cambridge English: First* (32), *Cambridge English: Advanced* (34) and *Cambridge English: Proficiency* (22). The first 1-minute long, mono-logic task of the test from each of the candidates' responses were extracted from the speech files. All candidate responses represented passing grades having received a standardised score of 75 or higher for the exam on a fixed scale out of 100. They represent 'average' learners at each level. Although the total length of some individual long-run responses (e.g. *Cambridge*

English: Proficiency) was 2 minutes or longer, only the first 1-minute response was included in the linguistic analysis for the consistency of the speech samples. For speech files which were shorter than 1 minute, normalisation (the number of linguistic features divided by the total number of words or the total number of the utterance time) took place before the final analysis. All 120 files were used for the lexical, grammatical analysis, and fluency analysis as part of discourse management, but only 115 files were used for pronunciation analysis because the other five files had speakers who exhibited creaky voice or vocal fry, which made the pitch extraction function of the PRAAT and the Computerized Speech Laboratory (CSL) software used in this study unable to read the data. Among the candidates, 76 were female and 44 were male. There were 21 first languages involved: 16 Spanish (including Mexican and Spanish), 11 Korean, 8 Italian, 7 Dutch, 6 French, 5 Chinese, 5 Russian, 4 Greek, 4 Portuguese, 4 Swedish, 3 German, 2 Swiss, 2 Japanese, 1 Brazilian, 1 Bulgarian, 1 Bolivian, 1 Austrian, 1 Turkish, 1 Arabic, 1 Colombian and 1 Estonian.

Linguistic analysis

Discourse management (fluency and coherence): In Cambridge English exams, management of discourse flow includes both fluency and coherence. As for fluency measures, the study examined speech rates (Kang 2008, Kormos and Denes 2004), and pause structures of the responses (Brown and Yule 1983, Kang 2008). Jamieson and Poonpon (2010) found that the number of key ideas and presence of an introduction were significantly related to score and task. Therefore, the study included fluency features as well as those two coherence features along with the use of conjunction devices such as transition or contrast. The following are the *discourse management* features included in the study: (a) syllables per second (the mean number of syllables produced per second); (b) mean length of run (the average number of syllables produced in utterances between pauses of 0.1 seconds and above); (c) number of silent pauses (the number of silent pauses per speech); (d) mean length of silent pauses; (e) number of hesitation markers (filled pauses); (f) mean length of filled pauses; (g) number of key ideas (i.e. the number of main subjects/topics) in each spoken text; (h) presence of introduction in each spoken text; and (i) the use of conjunction devices (i.e. addition, apposition, result, contrast and transition).

Lexical resource: The vocabulary size has been measured through vocabulary richness and type/token measures (Brown et al 2005). Vocabulary richness was calculated as a proportion of low and high-frequency vocabulary used in each spoken response. Vocabulary range was measured by type-token ratio. Iwashita et al (2008) found that increases in proficiency level were associated with an increase in the number of words produced (tokens) and a wider range of words (type). For lexical resources, the study included the following measures: (a) proportion of low and high-frequency vocabulary using the academic word list; (b) type-token ratio after measuring types and tokens; (c) lexical density (content words divided by the total number of words); (d) word families (number of words sharing the same origin); and (e) word length. The counts of the features were normalised to the first 100 words when needed.

Grammatical resource: Grammatical accuracy measures included global accuracy (Brown et al 2005) and specific types of errors (Iwashita et al 2008). Global accuracy means all errors produced are considered for analysis. These measures are suggested as possible predictors of oral language accuracy, according to empirical studies (Ortega 1999). The global accuracy varies significantly across proficiency levels (Iwashita et al 2008) and are task dependent (Jamieson and Poonpon 2010). The specific types of errors (tense marking, plural, preposition) were particularly significant features with high effect sizes that distinguish proficiency levels in spoken responses (Iwashita et al 2008). Global accuracy was examined by calculating error-free T-units as a percentage of the total number of T-units. Grammatical complexity was measured through verb–phrase ratio and occurrences of grammatical features. The number of verb phrases per T-unit (the verb–phrase ratio) was also identified as it was the most significant feature that distinguished proficiency levels among spoken responses (Iwashita et al 2008). A T-unit is defined as an independent clause and all its dependent clauses (Hunt 1970). In addition, grammatical complexity was examined by counting occurrences of prepositional phrases, passive structures, and adjectives as they revealed a significant effect on task types (i.e. independent vs. integrated task) and scores (Jamieson and Poonpon 2010). At the same time, data-driven grammatical complexity was examined by counting occurrences of the lexico-grammatical features generated by the Biber, Johansson, Leech, Conrad and Finnegan (1999) tagging programme.

The counts of the features were normalised to the first 100 words and only significant features will be discussed in this report. In the study, the following features were included for grammatical complexity and accuracy: (a) error-free T-unit; (b) total number of clauses; (c) T-unit complexity ratio; (d) total number of dependent clauses; (e) specific types of errors (e.g. tense marking, plural, preposition); and (f) occurrences of grammatical features (e.g. prepositional phrases, passive, adjectives, present tense).

Pronunciation: For measures of pronunciation, the project includes stress (Kang 2008, Kormos and Denes 2004) and pitch (Kang 2010, Pickering 2004). The variables selected were accented measures in pronunciation represented as 'acoustic fluency', the best predictor of rated oral performance (Kang, Rubin and Pickering 2010). Those variables have been also proven as pronunciation features that reveal high correlations between native speakers' (NSs') and non-native speakers' (NNSs') oral production in general (e.g. Kormos and Denes 2004, Pickering 2004). In addition, the study included nine tone choices (high-rising, mid-rising, low-rising, high-falling, mid-falling, low-falling, high-level, mid-level and low-level), as some of the tone choices (mid-falling and high-rising) are strong predictors of NNSs' oral proficiency (Kang et al 2010). The use of rising tone has been particularly emphasised in the native speaker's discourse context (Brazil 1997) as it can signal solidarity with speakers or common group or shared background. In the situation of discourse production, for example, it has been known that non-native, low-proficiency speakers tend to use low-falling tones between related propositions, whereas rising and mid-level tones would be anticipated by NS listeners (Wennerstrom

2000). Note that segmental features are excluded in this study due to low correlations found between segmental errors and listeners' judgements (e.g. Anderson-Hsieh, Johnson and Koehler 1992).

Therefore, the study included the following variables for pronunciation measures: (a) proportion of words with prominent stress (the proportion of prominent words to the total number of words); (b) number of prominent syllables per run; (c) overall pitch range (the pitch range of the sample based on the point of F0 minima and maxima appearing on prominent syllables per task); and nine tone choices (high-rising, high-level, high-falling, mid-rising, mid-level, mid-falling, low-rising, low-level and low-falling).

Data coding

The spoken responses were coded for linguistic features for each of the three scoring criteria (grammatical and lexical resources, discourse management and pronunciation). Coding was done both manually and automatically. All speech files were transcribed by two research assistants. Then, each of the linguistic features were detected by careful listening. Inter-coder reliabilities for all the linguistic analyses (interclass correlation coefficients) were .81 and higher. Note that although the analyses were conducted by using instruments and computer analysis software, analysts (or coders) had to reach agreement because instrumental analyses rely on accurate calibration of measuring devices and are often open to multiple interpretations (Crystal 2003). The transcripts were automatically tagged for grammatical features by Biber using his software (Biber et al 1999). For the measures of fluency and pronunciation features, speech samples were converted to digital wav files and transferred to the computer-assisted speech analysis programme, PRAAT (Boersma and Weenink 2007) for fluency and the CSL for intonation. For example, for fluency analaysis, features were identified and measured manually in milliseconds by using the waveform as shown in PRAAT. Transcripts were also used to analyse type-token ratio and vocabulary richness, using the web programme *VocabProfile* (Cobb 2002).

Statistical analysis

To determine the degree to which linguistic features distinguish CEFR speaking proficiency levels (B1–C2) in Cambridge English examinations, the data was analysed and interpreted through descriptive statistics and a series of ANOVAs. Thanks to the availability of information on candidates' proficiency level, *post hoc* analyses were conducted to characterise the type of patterns for each proficiency level.

Results

Salient linguistic features that distinguish CEFR speaking levels

The research question focused on identifying the overall salient linguistic features that distinguish CEFR speaking levels (B1–C2) in Cambridge English examinations for the following scoring criteria: (a) discourse management (coherence and fluency) (b) grammatical and lexical resources, and (c) pronunciation. For each scoring criterion, linguistic features were compared across four levels (*Cambridge English: Preliminary* (B1)*, Cambridge English: First* (B2)*, Cambridge English: Advanced* (C1) and *Cambridge English: Proficiency* (C2)).

Discourse management: Fluency

A series of ANOVAs indicated that most of the fluency variables (except for the number of filled pauses) were statistically significant across the levels of proficiency (see Table 1). In general, increases in proficiency level (from B1 to C2) were positively associated with increases in speech rate measures (syllable per second, mean length of run, and phonation time ratio). In terms of pauses and hesitation markers, as the proficiency increased, the pause frequency and length (number of silent pauses, mean length of silent pauses, number of filled pauses, and mean length of filled pauses) decreased from the lowest level to the highest, even though some variation appeared between adjacent levels.

Significant mean differences ($F_{3, 120} = 4.21$; $p < .007$, $d = .0.39$) were found in syllable per second between the B1 level and B2 and between B1 and C2. Respondents in *Cambridge English: First* and *Cambridge English: Proficiency* produced significantly more syllables in a given time than those in *Cambridge English: Preliminary*. No other group comparisons turned out to be significant. The mean length of run (utterances between major pauses of 0.1 seconds and above) progressively increased across the proficiency levels, even though the significant difference was found only in the comparison between B1 and C2. The same pattern was found in the phonation time ratio (the actual amount of speaking in a given time). The phonation time increased as proficiency increased.

Pauses and hesitation markers did support the findings

Table 1: Fluency features identified by proficiency levels

Fluency features	B1, *Preliminary* (N = 32)		B2, *First* (N = 32)		C1, *Advanced* (N = 34)		C2, *Proficiency* (N = 22)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Syllable per second*	1.86	0.38	3.05	2.88	2.64	0.09	3.11	0.08
Mean length of run*	3.13	0.84	4.18	3.05	4.73	1.30	7.90	4.05
Phonation time ratio*	0.67	0.08	0.70	0.07	0.72	0.09	0.76	0.06
Number of silent pauses*	31.29	6.95	39.01	8.79	32.11	7.84	32.63	7.21
Mean length of silent pauses*	0.69	0.21	0.48	0.13	0.57	0.17	0.36	0.13
Number of filled pauses	5.70	3.12	3.99	2.61	4.82	3.70	4.93	3.18
Mean length of filled pauses*	0.13	0.10	0.06	0.04	0.07	0.06	0.08	0.04

* represents variables that show statistical significance ($p < .05$) for overall analysis

Table 2: Coherence features identified by proficiency levels

Coherence features	B1, *Preliminary* (N = 32)		B2, *First* (N = 32)		C1, *Advanced* (N = 34)		C2, *Proficiency* (N = 22)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Number of key ideas*	4.44	0.94	3.41	0.61	3.32	0.76	2.91	0.98
Presence of introduction averaged per group	0.03	0.17	0.15	0.36	0.44	0.50	0.68	0.47
Conjunction device in total	5.40	2.68	7.81	3.09	8.73	3.25	10.23	3.01
Conjunction device: Addition	5.18	2.50	6.12	2.25	7.14	1.83	7.31	1.83
Conjunction device: Apposition*	0.00	0.00	0.00	0.00	0.09	0.28	0.50	0.59
Conjunction device: Result	0.18	0.39	1.03	1.57	0.94	1.41	1.45	1.50
Conjunction device: Contrast*	0.09	0.29	0.56	0.98	0.50	0.78	0.72	0.88
Conjunction device: Transition*	0.00	0.00	0.09	0.29	0.05	0.23	0.27	0.45

* represents variables that show statistical significance ($p < .05$) for overall analysis

that higher proficiency levels produced fewer disfluencies. Interestingly, B2 level respondents in the current speech samples produced significantly more silent pauses than the rest of the levels. However, the actual length of silent pauses in the B1 level was longer than any other level. Similarly, the mean length of filled pauses (um, ah, eh, etc.) were considerably longer in the B1 level, compared to the other higher level groups. Note that the number of filled pauses did not show any significant difference across levels.

Discourse management: Coherence

A series of one-way ANOVAs revealed significant differences among several coherence variables across the four levels (see Table 2). To be precise, the number of key ideas ($F_{3, 120} = 17.95$; $p = .000$, $d = 0.97$) and the use of some conjunction devices (i.e. apposition, contrast, and transition) ($F_{3, 120} > 15.60$; $p < .000$, $d > .84$) appeared differently across proficiency levels. The number of key ideas (i.e. the number of main subjects/topics), in particular, decreased substantially as the proficiency level increased. This means that compared to lower proficiency speakers, higher proficiency speakers may use longer or more complex sentences with multiple clauses to explain a relevant key topic so that, within a given time (1 minute), fewer topic changes may happen or irrelevant/new ideas may be less frequently introduced. That is, the same idea can be coherently described in a longer period of time. On the other hand, lower proficiency speakers may not have such ability that in a given time, they may more frequently change topics, introduce new subjects, or alter ideas.

The Tukey *post hoc* test showed that the difference was significantly strong in mean scores between *Cambridge English: Preliminary* and *Cambridge English: Proficiency* ($p < .05$). Other conjunction devices were generally used more frequently as candidates' proficiency went up. Significant differences were found in comparisons between *Cambridge English: Preliminary* and *Cambridge English: Proficiency* in the use of apposition, contrast and transition devices ($p < .05$). Candidates at the C2 level tended to use more transition devices than those at C1 ($p < .05$).

Grammatical and lexical resources

Thirty-one variables out of 125 grammatical features tagged by Biber et al's (1999) programme demonstrated significant mean differences across the four proficiency levels. Due to space constraints, only major findings are summarised in this article. These include: low-proficiency speakers tended to use more private words (e.g. I believe, I think, I feel, etc.), first person pronouns, third person pronouns, nouns and clausal coordinators (e.g. and, but, etc.) than other higher-level speakers. As proficiency increased, candidates used grammatically more complex and more structured expressions and phrases. The frequencies of the following features (i.e. the second person pronouns, emphatics, the pronoun 'it', 'be' as a main verb, subordinate clauses, perfect aspects, time adverbs, modals, conjunctions, etc.) also increased significantly with proficiency level, even with some inconsistency between adjacent levels.

The Tukey *post hoc* analysis results ($p < .05$) showed that the use of the pronoun 'it' and 'be' as a main verb was significantly different at each proficiency level; i.e. at each proficiency level, the frequency of using these words noticeably increased. While proficient speakers tended to utter certain causal adverbs (e.g. since or because) more frequently, they used other adverbial hedges (e.g. maybe), amplifiers (e.g. very), or clausal coordination (e.g. and) less habitually. It was also found that candidates at C2 used WH-clauses significantly more ($p < .05$) than those in other levels; no difference was revealed among candidates at B1, B2 and C1 in terms of this particular feature. When speaking, lower-proficiency speakers showed a tendency of using noun forms or short phrases more often than higher-proficiency speakers. As proficiency increased up to C2, the frequency of this noun usage seemed to have dropped. At the B1 level, there was no incident of using perfect aspect verbs. Then, as the proficiency increased, this feature appeared increasingly. The use of the 'to-infinitive' was another good indicator for distinguishing proficiency levels even though there was no strong difference found between B2 and C1. We could see a sizeable increase in the use of 'adverbial subordinator for condition' in the C2 level. This feature was hardly found in the B1 level.

Grammatical complexity and accuracy

The five grammatical complexity measures yielded positive results as seen in Table 3, which provides descriptive statistics. The expected gradient of increasing complexity per level ($F_{3, 120} > 2.74.21$; $p > .048$) was found for most of the measures (number of error-free T-units, total number

Table 3: Grammatical complexity features identified by proficiency levels

Grammatical complexity features	B1, *Preliminary* (N = 32)		B2, *First* (N = 32)		C1, *Advanced* (N = 34)		C2, *Proficiency* (N = 22)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
The total number of T-units	7.50	1.87	8.46	2.68	8.29	2.65	9.00	2.43
Number of error free T-units*	2.96	2.22	3.00	1.91	3.73	2.68	4.72	3.29
Total number of clauses*	11.34	3.84	16.18	4.78	17.73	5.45	25.00	5.58
T-unit complexity ratio*	1.51	0.34	2.01	0.70	2.19	0.53	2.89	0.65
Total number of dependent clauses*	3.53	2.68	7.37	4.53	8.97	4.29	15.36	5.52

* represents variables that show statistical significance ($p < .05$) for overall analysis

Table 4: Grammatical accuracy features identified by proficiency levels

Error types (frequency of errors)	B1, *Preliminary* (N = 32)		B2, *First* (N = 32)		C1, *Advanced* (N = 34)		C2, *Proficiency* (N = 22)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
article*	1.96	2.00	1.68	1.51	1.00	1.07	0.86	1.54
preposition*	1.71	1.22	1.55	1.05	1.18	1.05	0.65	0.82
adverb	0.28	0.52	0.25	0.50	0.29	0.52	0.28	0.56
pronoun	0.28	0.52	0.37	0.70	0.41	0.78	0.54	0.67
adjective	0.09	0.39	0.25	0.56	0.08	0.28	0.09	0.40
determiner	0.43	0.56	0.28	0.52	0.26	0.56	0.27	0.63
coordinator	0.25	0.50	0.43	0.61	0.41	0.92	0.54	0.73
subject	0.15	0.36	0.31	0.82	0.17	0.45	0.04	0.21
object	0.06	0.24	0.12	0.33	0.13	0.35	0.03	0.23
verb	0.25	0.50	0.09	0.29	0.20	0.47	0.31	0.48
copula	0.37	0.65	0.37	0.70	0.38	0.60	0.09	0.29
modal*	0.00	0.00	0.03	0.17	0.02	0.17	0.22	0.42
nominalization*	0.00	0.00	0.28	0.68	0.09	0.17	0.09	0.42
negator	0.00	0.00	0.06	0.24	0.02	0.17	0.00	0.00
comparative/superlative*	0.00	0.00	0.15	0.36	0.00	0.00	0.00	0.00
singular/plural*	0.65	0.86	0.81	0.99	0.35	0.54	0.35	0.72
tense	0.68	0.78	0.37	0.83	0.64	0.91	0.50	0.67
subject–verb agreement*	0.75	1.01	0.50	1.16	0.41	0.70	0.09	0.29
passive	0.09	0.53	0.03	0.17	0.14	0.43	0.13	0.35
relative clause	0.46	0.71	0.93	0.98	0.61	0.77	0.59	0.79
complement clause	0.09	0.39	0.09	0.39	0.17	0.45	0.04	0.21
non-finite clause	0.25	0.56	0.43	0.71	0.11	0.32	0.22	0.42
formation of subjunctive structure	0.03	0.17	0.03	0.17	0.02	0.17	0.09	0.29
formation of conditional structure*	0.00	0.00	0.03	0.17	0.23	0.49	0.31	0.38

* represents variables that show statistical significance ($p < .05$) for overall analysis

of clauses, T-unit complexity ratio, and total number of dependent clauses). The effect size varied, ranging from .06 (number of error-free T-units) to .56 (total number of dependent clauses). Note that the T-unit complexity ratio refers to the number of clauses per T-unit. The total number of clauses, T-unit complexity ratio, and the total number of dependent clauses were features that especially distinguish the proficiency levels of *Cambridge English: Preliminary, Cambridge English: First* and *Cambridge English: Proficiency*. These features of complexity were significantly different for each level ($p < .05$).

The descriptive statistics of seven grammatical accuracy variables (articles, prepositions, modals, normalisation, comparative/superlative, singular/plural, and subject–verb agreement) out of 24 measures revealed that *Cambridge English: Proficiency* is distinct from other levels, but the pattern is less clear at the adjacent levels (see Table 4), e.g. between B2 and C1. For several variables in particular, the frequency of errors tended to decrease as proficiency increased. For most of the features, the trend was not clearly linear but overall a gradual change was found. For example, grammatical errors in certain features increased from the lowest to the highest level; i.e. complicated features (e.g. tense, passive, relative clauses, complement clauses, non-finite clauses, formation of subjunctive structure, formation of conditional structure, pronoun, or coordinator) presented a very low error rate at *Cambridge English: Preliminary*, but at the higher levels the error rate increased conversely. (See the Discussion section for explanations for this phenomenon.)

Lexical analysis

A clear pattern was found in the lexical analysis of the given 120 speech files (see Table 5). Overall, most of the

Table 5: Lexical resource features identified by proficiency levels

Lexical features	B1, *Preliminary* (N = 32)		B2, *First* (N = 32)		C1, *Advanced* (N = 34)		C2, *Proficiency* (N = 22)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Types (range of words)*	44.16	8.96	60.06	10.92	65.15	12.55	81.82	9.83
Tokens (Number of words produced)*	83.38	21.49	125.84	25.66	130.85	31.03	174.36	22.73
Type-token ratio (TTR)*	0.54	0.08	0.48	0.05	0.51	0.07	0.47	0.04
K1 tokens*	66.34	19.32	108.22	25.68	112.88	25.99	150.68	24.20
K2 tokens	4.63	2.81	4.94	3.22	5.24	2.65	6.05	4.90
Academic word list (AWL) tokens*	0.63	1.1	1.25	1.24	1.79	1.95	3.09	1.57
Lexical density (content words/ total)*	0.49	0.05	0.48	0.05	0.46	0.05	0.43	0.04
Word families*	36.03	8.03	49.81	10.82	54.06	10.29	66.77	7.66
Word length*	3.71	0.22	3.90	0.24	3.91	0.32	4.30	0.27

* represents variables that show statistical significance ($p < .05$) for overall analysis
K1 = the most frequent 1,000 words of English
K2 = the second most frequent 1,000 words of English

Table 6: Pronunciation features identified by proficiency levels

Pronunciation features	B1, *Preliminary* (N = 32)		B2, *First* (N = 32)		C1, *Advanced* (N = 34)		C2, *Proficiency* (N = 22)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Number of prominent syllables per run (pace)	1.29	0.26	1.19	0.17	1.24	0.31	2.02	3.82
Proportion of prominent words to the total number of words (space)*	0.61	0.12	0.45	0.08	0.37	0.04	0.35	0.13
Overall pitch range*	67.90	38.16	75.59	28.62	93.26	42.24	103.08	38.67

* represents variables that show statistical significance ($p < .05$) for overall analysis

variables indicated significant differences across levels ($F_{3, 120} > 6.72.21$; $p > .000$). The effect size was .15 and higher for most of the significant variables. First of all, an increase in proficiency level was associated with an increase in the number of words produced (tokens) and a wider range of words (type). In addition, as a level changed from low to high, this increase in tokens and types was statistically significant ($p < .01$). C1 responses included significantly more tokens and types than those of B2 ($p < .05$). At the C2 level, these word ranges and amount were substantially greater than the rest of the levels. However, when the level changed from B2 to C1, these increases were only marginal and not significant. There was no consistent pattern found in the type-token ratio.

A similar pattern was found with candidates' use of the first 1,000 words (K1) and word families. The frequencies of the 1,000 word usage and the choices of various word families increased significantly as proficiency levels improved, i.e. from B1 to B2 and from C1 to C2. The changes from B2 to C1 were not statistically significant, however. No significance was found in the second 2,000 words (K2). Candidates' use of academic words demonstrated the same pattern as the ones above. Frequencies of academic word usage went up significantly and positively with level changes. Nevertheless, the increment from B2 to C1 did not reach any statistical significance. When it comes to lexical density, the results indicated that lower level respondents tended to produce more content words in relation to the total number of words produced (i.e. a gradual decrease from .49 down to .43). The significant difference of this content word ratio was found between B1 and C2. Finally, words chosen by high-proficiency candidates were longer than those by low-proficiency candidates even though a significant difference was found only in the comparisons of B1 and C2.

Pronunciation analysis

The first analysis of pronunciation included stress and overall pitch range measures (see Table 6). There were significant differences found in the proportion of prominent (stressed) words to the total number of words across levels. As proficiency increased, the proportion of stressed words decreased, especially in level changes from B1 to B2 and from B2 to C1. This means that low-proficiency speakers tended to place stress on words (regardless of their functions) more frequently than high-proficiency speakers. The overall pitch range was significantly different among the four groups of CEFR levels. Lower proficiency speakers in the levels of B1 and B2 had a more restricted pitch range than speakers who are in the advanced levels of C1 and C2. Changes from B2 to C1 and from C1 to C2 were statistically significant ($0 < .5$). No significant change was found in the number of prominent syllables.

In addition, the study examined nine tone choices which were compared across proficiency levels (see Table 7). Based upon the results of a series of one-way ANOVAs, high-rising, mid-rising, mid-falling and low-rising tones significantly differed by proficiency level ($F_{3, 120} > 3.37$; $p > .021$). The effect size was rather marginal, however, with .084 or a little higher. Mid-rising and high-rising tones were prominently and positively associated with proficiency; i.e. the use of mid- and high-rising tones increased substantially as levels went up. On the other hand, falling or level tone choices were more frequently used by low-proficiency speakers.

Table 7: Tone choices identified by proficiency levels

Tone choices	*p*	B1, *Preliminary* (N = 31)		B2, *First* (N = 31)		C1, *Advanced* (N = 31)		C2, *Proficiency* (N = 21)	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
High-rising*	0.01	4.42	5.72	6.13	6.60	9.32	11.50	12.14	13.23
High-level	0.14	1.24	1.99	1.23	2.35	0.25	1.04	0.76	2.11
High-falling	0.62	7.92	7.14	9.91	7.55	9.03	9.25	10.89	10.00
Mid-rising*	0.00	24.69	11.45	29.81	11.83	35.29	17.16	45.30	14.00
Mid-level	0.06	9.36	12.16	8.42	9.81	5.22	7.71	3.64	5.54
Mid-falling*	0.00	50.64	12.78	40.52	15.32	27.77	12.44	28.71	15.28
Low-rising*	0.04	5.17	9.07	2.09	3.88	2.85	6.06	0.42	1.07
Low-level	.15	2.05	5.89	0.76	1.69	0.08	0.49	0.22	1.54
Low-falling	.75	1.71	3.05	1.81	4.22	2.09	3.89	0.99	2.00

* represents variables that show statistical significance ($p < .05$) for overall analysis

Discussion

The essence of the criterial feature concept in the CEFR (Hawkins and Filipović 2012) is that if we make the distinguishing properties explicit at each linguistic level (grammatical and lexical, phonological and semantic), we can detect a set of linguistic features which can distinctively specify the CEFR's descriptors for each proficiency level. It hypothesises that there are certain linguistic properties that are characteristic and indicative of L2 proficiency at each level (Hawkins and Filipović 2012). Accordingly, the current study has attempted to illustrate linguistic features in speaking performances that can distinguish the different CEFR levels.

Overall findings suggest that there are distinctive differences in linguistic features across CEFR speaking levels (*Cambridge English: Preliminary* (B1), *Cambridge English: First* (B2), *Cambridge English: Advanced* (C1) and *Cambridge English: Proficiency* (C2)). For fluency features as part of discourse management, the increases in proficiency level (from B1 to C2) were positively associated with increases in speech rate measures (syllable per second, mean length of run and phonation time ratio). At the same time, the pause frequency and length (number of silent pauses, mean length of silent pauses, number of filled pauses, and mean length of filled pauses) were negatively associated with proficiency level. In other words, proficient candidates produced more syllables per second, and longer utterances between major pauses, and used fewer silent pauses and hesitation markers. Note that the number of filled pauses did not show a linear trend across the proficiency levels. As de Jong, Groenhout, Schoonen and Hulstijn's (2012) findings suggest, there might be an interaction between the use of filled pauses and that of silent pauses, i.e. L2 speakers who tend to use many filled pauses and repetitions may use shorter pauses. In addition, while the number of key ideas substantially dropped as proficiency increased, certain conjunction devices (i.e. apposition, contrast and transition) were more frequently found among higher-proficiency speakers than lower-proficiency speakers. This finding is understandable particularly because less competent speakers may use more content words instead of complete, complex sentences. In terms of level differences, although there were clear distinctions between B1 and C2 with regard to candidates' speaking characteristics, linguistic features were less distinctively found in adjacent levels (e.g. between B2 and C1).

By using Biber et al's (1999) lexico-grammatical tagging programme, the study found distinctive patterns in various grammatical items. Thirty-one grammatical features were identified as strong indicators of distinguishing proficiency levels. Compared to high-level respondents, low-proficiency speakers used more private words, first person pronouns, nouns and clausal coordinators. Not surprisingly, high-level candidates used grammatically more complex and more structured expressions and phrases such as emphatics, 'be' as a main verb, subordinate clauses, perfect aspects, time adverbs, modals, conjunctions, etc. They used more complicated adverbial expressions such as causal adverbs (e.g. since), WH-clauses, T-infinitive, adverbial subordinator for condition, but used less simplified ones, for instance, adverbial hedges (e.g. maybe), amplifiers (e.g. very), or clausal coordination (e.g. and). Lower-proficiency speakers tended to use noun forms or short phrases more frequently than higher-proficiency speakers. These results concur with the findings of previous studies investigating written texts: advanced learners used more features such as be-copula as the main verb (Hinkel 2003) or pronoun, hedges, verbs, WH-clause, subordinators (Espada-Gustilo 2011, Grant and Ginther 2000). Overall, the occurrence of grammatical features which involve more complicated forms of structure or arrangement seem to increase as proficiency levels increased but simplified forms or content-word based formations such as nouns were used among low-proficiency speakers.

The expected gradient of increasing complexity per level was found for most of the measures (number of error-free T-units, total number of clauses, T-unit complexity ratio and total number of dependent clauses). These findings are in line with the literature (e.g. Iwashita et al 2008), using the iBT TOEFL spoken data. The complexity of these features was significantly different for each level ($p < .05$). Moreover, in terms of grammatical accuracy, the frequency of errors decreased as proficiency increased. Especially, variables such as articles, prepositions, modals, normalisation, comparative/superlative, singular/plural, and subject–verb agreement in the C2 level were distinctively different from those in other levels. One interesting thing to note is that grammatical errors in certain features (e.g. passive, relative clauses, complement clauses, non-finite clauses, formation of subjunctive structure)

increased with proficiency. It can be speculated that these complex features were not found at the lower level as candidates might avoid using them or not hold this ability. As their command of English improved, they attempted to use them more and consequently made more mistakes. In other words, this is the tension between complexity and accuracy. As candidates tend to take more risks when speaking, their accuracy may go down.

The results of the lexical analysis also revealed noticeable patterns across four different levels. Increase in proficiency resulted in an increase in the number of words produced (tokens) and a wider range of words (types). Interestingly, each level presented salient features that distinguished its level from others such as B1 vs B2 or C1 vs C2. This result is in line with recent findings of other studies (e.g. Galaczi and ffrench 2011, Iwashita et al 2008) that as examinees develop in proficiency, they produce more words (tokens) and display a wider range of vocabulary by using different words (types). There was also a significant increase in the frequencies of the 1,000-word usage and the choices of various word families, and academic word usage as proficiency levels improved. In addition, words chosen by high-proficiency candidates were longer than those by low-proficiency candidates, especially in the level changes from B1 to C2. However, distinctiveness of lexical features between the B2 and C1 level was less obvious.

Finally, in terms of pronunciation features, low-proficiency speakers emphasised words with stress more frequently than high-proficiency speakers. Typically, low-proficiency NNSs use primary stress on every lexical item, regardless of its function or semantic importance (Kang 2010, Wennerstrom 2000). In addition, lower-proficiency speakers in the B1 and B2 levels had a more restricted pitch range than speakers in C1 and C2. Given that the importance of intonation and tone use is well recognised (Derwing and Munro 2005, Kang et al 2010), the current study also included nine tone choices (high-rising, mid-rising, low-rising, high-level, mid-level, low-level, high-falling, mid-falling, and low-falling) for pronunciation measures. Among these nine tone choices, the frequencies of high-rising, mid-rising, mid-falling and low-rising tones were statistically different across proficiency levels. The findings were parallel with Kang et al's study (2010). That is, while mid-rising and high-rising tones were positively associated with proficiency, mid-level and low-falling tones were negatively associated with proficiency. In other words, candidates in C2 used a wide range of tone choices including native-like rising tones, whereas those at B1 and B2 chose tones limitedly. Overall, these tone choice variables appeared to be good indicators for distinguishing candidates' speaking performance across CEFR levels for the criterion of pronunciation.

In sum, distinctive patterns in linguistic features across CEFR speaking levels (*Cambridge English: Preliminary* (B1) *Cambridge English: First* (B2), *Cambridge English: Advanced* (C1) and *Cambridge English: Proficiency* (C2)) were found, even though there was some fuzziness of distinctions at adjacent levels. The complexity of the configuration of components in any overall judgement of proficiency, and the fuzziness of distinctions between levels seem to be unavoidable (Lee and Schallert 1997). As Douglas and Selinker (1993) argued, it is possible that speakers may produce qualitatively quite different performances and yet receive similar ratings. Still in this study, many of the linguistic features in each category were linked to a proficiency level, which can be relatively distinguished from others. It is hoped that the findings of the current project can be used as resources for the empirical validity evidence for the Cambridge English Speaking tests at CEFR levels B1 to C2.

Conclusion

The study sought to identify linguistic features that distinguish levels of candidates' performances in one suite of high-stakes speaking tests, i.e. the Cambridge English examinations. The outcomes of the study have made two issues explicit: (1) there are salient linguistic features useful for distinguishing scoring criteria and tentatively defining certain criterial features at each level; and (2) there are objective (not impressionistic) differences between high-scoring performances and low-scoring performances. The findings offer implications for enhanced scoring criteria, rater development and English language pedagogy. That is, specific linguistic features and their contribution to each proficiency level can be integrated into scoring descriptors in the Cambridge English Speaking tests and further inform the future development of automated scoring systems. Moreover, salient features identified in this study (e.g. speech rate or the number of key ideas for discourse management; pitch range or number of stressed words for pronunciation; grammatical accuracy for grammatical resource; or a wide range of vocabularies for lexical resource) can inform rater training of L2 speaking assessment. As discrete linguistic features of candidates' oral performance can facilitate the process of finding scoring benchmarks, raters can be trained to pay special attention to those characteristics. Finally, when NNSs' specific linguistic features for each proficiency level are documented, concrete advice can be given to English as a Second Language/English as a Foreign Language teachers so that students can better utilise their linguistic repertoires in high-stakes test situations.

Note that the current study analysed speech data from one mono-logic speaking task. That is, it did not examine the effect of task difference, which can potentially affect the distribution patterns of features across proficiency levels. Interpretation of results should be modified when applying the linguistic patterns examined in this study to other speaking tasks such as paired or group interactions. In addition, linguistic measures included are comprehensive, yet not exhaustive, with regard to assessing L2 oral proficiency. Future research can further investigate other linguistic components such as content-related discourse features or segmental aspects of pronunciation in oral assessment.

References

Anderson-Hsieh, J, Johnson, R and Koehler, K (1992) The relationship between native speaker judgments of non-native pronunciation and deviance in segmentals, prosody and syllable structure, *Language Learning* 42, 529–555.

Biber, D, Johansson, S, Leech, G, Conrad, S and Finnegan, E (1999) *Longman Grammar of Spoken and Written English*, Essex: Pearson Education Limited.

Boersma, P and Weenink, D (2007) *Praat*, www.praat.org (Version 4.5.25).

Brazil, D (1997) *The Communicative Value of Intonation in English*, Cambridge: Cambridge University Press.

Brown, A, Iwashita, N and McNamara, T (2005) *An Examination of Rater Orientations and Test-taker Performance on English-for-Academic-Purposes Speaking Tasks*, TOEFL Monograph Series MS-29, Princeton: Educational Testing Service.

Brown, G T L and Yule, G (1983) *Teaching the Spoken Language: An Approach Based on the Analysis of Conversational English*, Cambridge: Cambridge University Press.

Chapelle, C, Enright, M and Jamieson, J (Eds) (2008) *Building a Validity Argument for TOEFL*, New York: Routledge/Taylor and Francis Group.

Cobb, T (2002) *The Web Vocabulary Profile*, available online: www.lextutor.ca/vp/

Crystal, D (2003) *English as a Global Language*, Cambridge: Cambridge University Press.

Davies, A (2008) *Assessing Academic English: Testing English Proficiency 1950–1989 – The IELTS Solution*, Studies in Language Testing volume 23, Cambridge: UCLES/Cambridge University Press.

de Jong, N, Groenhout, R, Schoonen, R and Hulstijn, J H (2012) L2 fluency: Speaking style or proficiency? Correcting measures of L2 fluency for L1 behavior, *Applied Psycholinguistics*, available online: www.academia.edu/2240591/L2_fluency_speaking_style_or_proficiency_Correcting_measures_of_L2_fluency_for_L1_behavior

Derwing, T and Munro, M (2005) Second language accent and pronunciation teaching: A research-based approach, *TESOL Quarterly* 39, 379–397.

Douglas, D and Selinker, L (1993) Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants, in Douglas, D and Chapelle, C (Eds) *A New Decade of Language Testing Research*, Alexandria: TESOL Publications, 235–256.

Espada-Gustilo, L (2011) Linguistic features that impact essay scores: a corpus linguistic analysis of ESL writing in three proficiency levels, *The Southeast Asian Journal of English Language Studies* 17 (1), 55–64.

Field, J (2011) Cognitive validity, in Taylor, L (Ed.) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 65–111.

Galaczi, E D and ffrench, A (2011) Context validity, in Taylor, L (Ed.) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 112–170.

Garman, M (1990) *Psycholinguistics*, Cambridge: Cambridge University Press.

Grant, L and Ginther, A (2000) Using computer-tagged linguistic features to describe L2 writing differences, *Journal of Second Language Writing* 9 (2), 123–145.

Hawkins, J A and Filipović, L (2012) *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*, English Profile Studies volume 1, Cambridge: UCLES/Cambridge University Press.

Hinkel, E (2003) Simplicity without elegance: Features of sentences in L1 and L2 academic texts, *TESOL Quarterly* 37 (2), 275–301.

Hunt, K W (1970) Syntactic maturity in school children and adults, *Monographs of the Society for Research in Child Development* 35, 1–67.

Iwashita, N, Brown, A, McNamara, T and O'Hagan, S (2008) Assessed levels of second language speaking proficiency: How difficult? *Applied Linguistics* 29, 24–49.

Jamieson, J and Poonpon, K (2010) *Rating guide development for speaking's dimensions: Delivery, language use, topic development*, paper presented at Language Testing Research Colloquium, Cambridge.

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

Kang, O (2008) Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness, *Spaan Fellow Working Papers* 6, 181–205.

Kang, O (2010) Relative salience of suprasegmental features on judgements of L2 comprehensibility and accentedness, *System* 38, 301–315.

Kang, O, Rubin, D and Pickering, L (2010) Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English, *Modern Language Journal* 94, 554–566.

Kormos, J and Denes, M (2004) Exploring measures and perceptions of fluency in the speech of second language Learners, *System* 32, 145–164.

Lee, J W and Schallert, D (1997) The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL Context, *TESOL Quarterly* 31, 713–739.

Ortega, L (1999) Planning and focus on form in L2 oral performance, *Studies in Second Language Acquisition* 21, 109–48.

Pickering, L (2004) The structure and function of intonational paragraphs in native and non-native speaker instructional discourse, *English for Specific Purposes* 23, 19–43.

Taylor, L (2001) Revising the IELTS Speaking test: Developments in test format and task design, *Research Notes* 5, 3–5.

Taylor, L (2011) Introduction, in Taylor, L (Ed.) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 1–35.

Wennerstrom, A (2000) The role of intonation in second language fluency, in Riggenbach, H (Ed.) *Perspectives on Fluency*, Ann Arbor: University of Michigan, 102–127.

ALTE PARIS 2014

ALTE 5th International Conference

10–11 April 2014, Paris, France

Language Assessment for Multilingualism: promoting linguistic diversity and intercultural communication

Call for Papers open

ALTE (the Association of Language Testers in Europe) invites you to submit a paper for the ALTE 5th International Conference to be held in Paris from 10–11 April 2014.

Papers are welcomed in English, French, German, Italian and Spanish, and all proposals must relate to the conference theme: *Language Assessment for Multilingualism: promoting linguistic diversity and intercultural communication*.

The deadline for the submission of papers is **30 September 2013**

Visit **www.alte.org/2014** for more information.

0925954657

To subscribe to *Research Notes* and download previous issues, please visit:
www.cambridgeenglish.org/research-notes

Contents:

For further information visit the website:
www.cambridgeenglish.org

Cambridge English
Language Assessment
1 Hills Road
Cambridge
CB1 2EU
United Kingdom
Tel. +44 1223 553997

ALTE Association of Language Testers in Europe

A DIVISION OF CAMBRIDGE ASSESSMENT