

LANGUAGE TESTING FORUM 2012

Graduate School of Education



16-18 November
www.bristol.ac.uk/education/ltf2012



Sponsors



Technical communication

Web and email access will be available in all conference and exhibition areas via the **VisitorNet** Wi-Fi service:

wireless.bristol.ac.uk/visitornet

VisitorNet passwords and connection instructions will be available for delegates to collect from the registration desk.



Fast **eduroam** wireless Internet is also available for delegates from organisations participating in the eduroam federation.

eduroam users will be able to connect immediately using the username and password they already have from their home institution (eduroam users should have set-up their computer at their home organisation before coming to Bristol).

PROGRAMME

Friday 16 November 2012		
17:00-18:00	Registration (Sponsorship from British Council for student registrations)	4 th Floor Foyer, 35 Berkeley Square
18:05-18:15	Welcome and Introduction	
18:15-19:15	Opening Lecture Professor Patricia Broadfoot University of Bristol	The next frontier in educational assessment: what can technology-enhanced approaches offer?
19:15-20:15	Drinks Reception (Subsidized by Oxford University Press)	4 th Floor Foyer, 35 Berkeley Square
20:30-23:00	Dinner (Sponsorship from Cambridge ESOL for student places)	Jamie's, Park Street
Saturday 17 November 2012		
09:00-09:30	Paper 1 Tony Green	Promoting assessment literacy in the Russian Federation: the ProSET project
09:30-10:00	Paper 2 Hanan Khalifa, Christine Walker	An investigation into intensive language provision and external assessment in primary education in Vietnam
10:00-10:30	Paper 3 Angeliki Salamoura, Coreen Docherty, Miranda Hamilton	Using external language assessment in a primary and secondary school context: The learners' perspective
10:30-11:00	Tea/Coffee Break	
11:00-11:30	Paper 4 Zahra Al Lawati	Investigating characteristics of test specifications and item writer guidelines, and their effect on item development
11:30-12:00	Paper 5 Claudia Harsch, Johannes Hartig	Effects of selected item characteristics on difficulty and dimensionality of reading and listening comprehension tests
12:00-12:30	Paper 6 Glyn Jones	Test item order: what difference does it make?
12:30-14:00	Poster oral presentations (5 mins x 9), see the running order on p.3 Lunch (subsidized by Trinity College London)	4 th Floor Foyer, 35 Berkeley Square

14:00-14:30	Paper 7 Laura Vilkaite, Henrik Gyllstad Norbert Schmitt	Validation evidence for and against the 'Vocabulary Size Test'
14:30-15:00	Paper 8 Ying Zheng, Li Li	Test takers, test trainers, and test developers: Are they speaking the same language?
15:00-15:30	Paper 9 Yang Lu	Exploring the criterion-related validity of New HSK for a U.K. university's Mandarin Chinese courses
15:30-16:00	Tea/Coffee Break	
16:00-16:30	Paper 10 Abdul Halim Abdul Raof, Anie Attan, Noor Abidah Mohd. Omar, Masputeriah Hamzah, Azian Ab. Aziz	Assessing graduating students' communication skills: Issues and challenges
16:30-17:00	Paper 11 Vanda Papafilippou	Discipline, punish and authority: the power of English language tests within the Greek context
17:00-17:50	Drinks Reception (Subsidized by British Council)	4 th Floor Foyer, 35 Berkeley Square
18:00-20:30	Dinner	Jamie's, Park Street
Sunday 18 November 2012		
09:00-09:30	Paper 12 John Field	Cognitive validity in language testing: theory and practice
09:30-10:00	Paper 13 Ardeshir Geranpayeh, Mark Elliot	Validation of L2 listening comprehension tests within a socio-cognitive framework
10:00-10:30	Paper 14 Fumiyo Nakatsuhara, John Field	Examiner interventions in oral interview tests: what are the listening demands they make upon candidates?
10:30-11:15	Tea/Coffee Break, Group Photo (Brandon Hill, weather permitting)	
11:15-11:45	Paper 15 Talia Isaacs	Interactional patterns and interlocutors' perspectives on a paired speaking test task
11:45-12:15	Paper 16 Gad Lim, Evelina Galaczi, Nahal Khabbazzbashi	Descriptor salience and clarity in rating scale development and evaluation
12:15-13:15	Closing Lecture Professor Barry O'Sullivan British Council	Standards and assessment
13:15-13:25	Farewell and LTF 2013	
13:25-14:25	Lunch (Subsidized by Pearson)	4 th Floor Foyer, 35 Berkeley Square

Posters

(See pages 18-23 for the abstracts)

The relationship between test-takers' first language, listening proficiency and their performance on paired speaking tests

Suwimol Jaiyote University of Bedfordshire

Teachers' perspective of washback effects of English reading tests in Japanese university entrance examinations

Yumiko Moore University of Exeter

An investigation into the construct validity of an academic writing test in English with special reference to the Academic Writing module of the IELTS test

Ahmad Alsagoafi University of Exeter

Korean teachers' attitudes toward the National English Ability Test in Korea: Implications for washback

Kyeonghwa Lee University of Bristol

Suitability of Pearson PTE Academic as an exit test for a course of academic English

Nathaniel Owen University of Leicester

The English test effect of the Thai university entrance examination on English teachers and students

Chamnan Para University of Bristol

Automated assessment of lexical diversity and Ngrams in essays at different levels of the CEFR

Jeanine Treffers-Daller, Shirley Williams, Patrick Parslow, University of Reading

Medical student's perceptions of fairness in written exams

Enas Elmansuri University of Bristol

A Thinkpiece - ESL curriculum and testing for vocational purposes: Is it time to bring it all together?

Maria Doyle SELECT Consultants, Australia

ABSTRACTS

INVITED LECTURES

Opening lecture on Friday 16 November 2012

The next frontier in educational assessment: what can technology-enhanced approaches offer?

Patricia Broadfoot, University of Bristol



The contemporary world is increasingly dominated by interactive technologies and the ubiquitous use of the Internet. Yet current assessment approaches remain curiously untouched by such developments. In this presentation, I will briefly describe some of the cutting edge developments in this respect and discuss how far new technologies may usefully be harnessed in the service of better, more efficient assessment but how far too, such technologies may bring with them hidden dangers.

Closing lecture on Sunday 18 November 2012

Standards and assessment

Barry O'Sullivan, British Council/Roehampton University, London



Before we can consider either setting, maintaining or applying standards we must first be clear what we mean by the term 'standards'. Standards can be considered from a number of different perspectives: standards as a set of guidelines, standards as quality and standards as level. The first of these definitions refers to those procedures identified as being necessary in order to develop appropriate tests. The second is often perceived as a quality associated with an entity ('the standard of English is falling'), while the latter concept refers to educational performance standards where standards represent an attempt to define the progressive stages of knowledge and skills that students should possess at key points in their learning.

In this talk, I will focus on the latter two understandings of standards, considering the key role of assessment in the setting and maintaining of language standards. I will argue that the success of any learning system hinges on the quality of the original performance standards, the manner in which the standards are delivered in the classroom and the degree to which assessment is integrated into the whole system.

ABSTRACTS OF PAPER PRESENTATIONS

Saturday 17 November 2012

09:00-09:30 Paper 1
Tony Green

**Promoting assessment literacy in
the Russian Federation: the ProSET
project**

This paper reports on the progress of a project to develop introductory teacher training course modules on language assessment for language teachers in Russian secondary education. Currently in an initial pilot phase, the modules will be integrated into the Russian federal university curriculum for trainee language teachers. The modules include lecture and workshop elements and are supported by new materials for students and instructors. Additional online distance variations will also be developed in order to facilitate access to training for in-service teachers. The project will set up and equip regional Professional Development centres on Testing and Assessment intended to sustain assessment expertise in partner universities. Through a process of capacity building in language testing and assessment, the project will contribute to improving assessment literacy among school teachers and the wider constituency. The project will impact both the programmes offered by universities and the wider practice of teaching and assessment in schools across the Russian Federation.

09:30-10:00 Paper 2
Hanan Khalifa, Christine Walker

**An investigation into intensive
language provision and external
assessment in primary education in
Vietnam**

The Ministry of Education and Training in Vietnam is working with Cambridge ESOL to fulfil its strategic objective to raise English Language learning standards by 2020. This presentation focuses on the implementation of the Ministry's objective in Ho Chi Minh City through two interventions, an Intensive English Programme and a standardised external assessment. We first describe the impact study in which data were collected in a mixed method research design. Qualitative data collection instruments comprised interviews with policy makers, district heads, school principals and student focus groups in 24 schools. Quantitative analyses were also undertaken on teacher and parent surveys (113 teachers and 2683 parents across 59 schools) and test score data. The surveys sought teacher and parental views on learning English, assessment in general and on Cambridge English: Young Learners exams in particular; perceptions of the Ministry's intervention and learner motivation and progression, together with collecting information about classroom practices.

Test results in two consecutive academic years (2010-11 and 2011-12) were analysed to ascertain whether standards of English had improved or not. Next we summarise the results from the study, which suggest a generally positive attitude towards learning English, of assessment in general and of the specific external assessments used in this context. Whilst learner motivation was shown to be quite high, the results also suggested that test anxiety was salient to parents and teachers but not to learners. The study also revealed that learners progressed in their language proficiency as a result of the intensive programme and external assessment, with the most improvement in speaking; there was also evidence of positive changes in teaching practice such as increased adoption of Assessment for Learning (AfL) principles. We conclude this presentation by reporting some intended and unexpected effects of this programme for the stakeholders involved and recommendations for further study.

10:00-10:30 Paper 3 Angeliki Salamoura, Coreen Docherty, Miranda Hamilton	Using external language assessment in a primary and secondary school context: The learners' perspective
---	--

Each year educational initiatives are put in place to try to raise English language standards in schools and it is increasingly common for standardised language assessments to be introduced as part of these initiatives. It is now widely understood that tests can have multi-faceted, complex effects on learners. For example, research shows that learners' approach and commitment to learning and test preparation is most intense when they perceive the status of the test as high (Stoneman 2006) and the results as important (Green 2007). Researching learners' views of the tests they take is therefore crucial to ensure that they are having a positive impact on learning. This paper investigates learners' perceptions of external standardised examinations in a bilingual programme introduced by an association of primary and secondary schools in Spain. The external examinations are a key component of the programme in a school system where such examinations are not typical. To what extent are learners more motivated to learn and use English or more anxious? In what ways do younger learners' perceptions differ from those of older learners? A mixed methods research design was used. Questionnaire data were collected from 291 students in 20 schools and analysed quantitatively. A subsection of younger learners (103 students from 13 schools; 9-12 year olds) also participated in focus groups and 'draw-a-picture' sessions, an alternative methodology used to elicit opinions in a pictorial form that suits younger learners' cognitive preferences. The focus group data were analysed using inductive theme analysis (Patton 2002) to identify major trends, whereas the 'draw-a-picture' data were explored using content analysis (Harris et al. 2009). Test data were reviewed to examine progress before all data were triangulated. The findings will be discussed with reference to the programme's objectives and recent research on learners' views of assessment from diverse school settings (e.g. Hong Kong's examination-oriented system; Carless and Lam 2012) to draw implications for impact models in language testing.

10:30-11:00

Tea/Coffee Break

11:00-11:30 Paper 4
Zahra Al Lawati

Investigating characteristics of test specifications and item writer guidelines, and their effect on item development

Test specifications are defined in the language testing literature as generative blueprints for test design (Alderson et al., 1995; Carroll, 1980; Davidson Lynch, 2002 Popham, 1978; Tinkelman, 1971), and it is emphasised that test specifications should be provided to item writers to help them produce the intended items. For the same purpose item writer guidelines are developed. Surprisingly, not much research has been done on test specifications and item writer guidelines. As a consequence, very little is known about the development process of these two documents, their use, and how they lead to better item writing. Similarly, not much is known about the characteristics of the two documents, the exact relationship between the two documents, and their relationship with the developed items. The study reported on in this paper investigates the nature of test specifications and item writer guidelines, and the effect of their use on ESL reading item development. It also aims to find out to what extent the language proficiency level descriptors of the Common European Framework (CEFR) lend themselves to producing clear test specifications and item writer guidelines. In addition, the amount of details that should be provided in test specifications and item writer guidelines in order to help item writers produce the intended items is investigated, as well as how the two documents can predict the level of difficulty of the items produced. The research aims and design will be described, and an overview will be given of the qualitative and quantitative approaches used to study the two documents. In particular, interview data collected from developers of test specifications and item writer guidelines, and focus group and interview data from item writers will be discussed.

11:30-12:00 Paper 5
Claudia Harsch, Johannes Hartig

Effects of selected item characteristics on difficulty and dimensionality of reading and listening comprehension tests

Research on central topics in language testing such as construct validity, item difficulties, or dimensionalities of latent abilities can be facilitated by employing item characteristics. The project we present uses parallel listening and reading tasks in English as a foreign language (EFL) to investigate the influence of specific item characteristics on item difficulties (Freedle/Kostin, 1993; Gorin/Embretson, 2006) and to validate assumptions about the test constructs (Embretson, 1998). More specifically, we set out to investigate the dimensionalities of reading and listening comprehension, as well as the influence of task types and related cognitive operations on item difficulty. The following characteristics were selected on the basis of existing models in reading and listening comprehension (Buck, 2001; Khalifa & Weir, 2009), and on our prior data analyses of EFL tests: cognitive operations (based on models on listening and reading processes), text difficulty (linguistic level), speech rate and item format (MC and short open answers). Six tasks were developed encompassing 10-15 items each, using the same texts/inputs and items for assessing both listening and reading comprehension. This resulted in a matrix of 164 items systematically varied over

characteristics and skills. The tests were piloted in 2010 (n=100 per item) and, after revision, administered to over 2200 15-year-old students in a multi-matrix design in early 2011, resulting in about 1100 answers per item. The test data were analysed using explanatory multidimensional Rasch-Models (De Boeck/Wilson, 2004; Wilson/De Boeck/Carstensen, 2008). These models allow the incorporation of a-priori item characteristics into the measurement model and they allow for testing the hypothesized effects of item characteristics on item difficulties. The results can be used to examine theory-based validity (Weir, 2005) and to develop item writer guidelines. Furthermore, they can inform decisions on score reporting regarding dimensionality and on setting cut-scores when it comes to reporting reading and listening comprehension.

12:00-12:30 Paper 6
Glyn Jones

**Test item order: what difference
does it make?**

Studies of the effect of test item order on test taker performance have generally set out to test one or other of two slightly different types of hypothesis, which can be characterised broadly as follows:

- Item difficulty is influenced by the position of an item in the sequence: the later the easier. If supported, this hypothesis may constitute evidence of a practice effect; test takers get better at answering items as they get more accustomed to the required cognitive process and manipulations
- Item difficulty is influenced by the difficulty of immediately preceding items: the more difficult the preceding items, the more difficult the item is (regardless of its ordinal position). If supported, this may constitute evidence of an anxiety effect; being confronted with a difficult item induces stress, adversely affecting performance on subsequent items.

In the present study is based on scores from a large number (N>10000) of candidates who participated in trials of a high stakes English Language proficiency test. Several parallel test forms were compiled for the trial such that

- the same item can occur in different ordinal positions in different test forms, and
- the same item can occur in the same ordinal position in different test forms, but preceded by items of different degrees of difficulty.

By analysing a large number of instances of both of these conditions the study endeavours to ascertain the extent to which either or both of the above effects - a practice effect and/or an anxiety effect - may be adduced in the case of this test.

**12:30-14:00 Poster oral presentations (5 mins
 each),**

**see the running order on p.3, and
the abstracts on pp.18-23**

Lunch
(subsidized by Trinity College
London)

4th Floor Foyer, 35 Berkeley Square

14:00-14:30 Paper 7

**Laura Vilkaite, Henrik Gyllstad
Norbert Schmitt**

**Validation evidence for and against
the 'Vocabulary Size Test'**

The Vocabulary Size Test (VST) (Nation and Beglar, 2007) is increasingly being used for both research and pedagogy, despite the fact that there is still little evidence supporting its validity for these purposes. This study investigates a number of issues particular to the VST. L2 participants in Lithuania, Sweden, and the UK were given the original ten VST items from one of three one-thousand-word frequency levels (3,000, 6,000, or 9,000). They were also given another 20 newly-created VST items from the level. They were then interviewed one-on-one about their knowledge of these words, plus an additional 70 from the level (total 100). Thus, their results at various sampling rates (10/1,000 to 30/1,000) could be compared to the in-depth interview which sampled at a 100/1,000 rate. During the interview, the participants were also queried about their test-taking behaviour, guessing, and strategic approach to the VST. Sample Results: - The 10-item test has an insufficient sampling rate, correlating at only .80 with the interview measure. A 30-item test would be much more satisfactory, with a .93 correlation. - Guessing is a serious threat to the validity of the VST. 57% of the participants were able to guess 2 or more of the original 10 items, and 30% were able to guess 3+. Given that each item represents 100 word families, this can lead to very substantial overestimation of vocabulary size. This guessing effect was particularly problematic with less-proficient learners, with overestimates of up to 50% for some participants. - The guessing behavior was largely strategic, sometimes aided by weakly-written items. Six major strategies were identified. - Cognates were largely facilitative for participants who knew at least 1/3 of the words in a level, but learners with lower vocabulary sizes had difficulty utilizing cognates in the test.

14:30-15:00 Paper 8

Ying Zheng, Li Li

**Test takers, test trainers, and test
developers: Are they speaking the
same language?**

Literature shows that sometimes what test developers think they are testing may not be the same as what test takers think the test is measuring (Fox, Cheng, & Zheng, 2007). Likewise, test takers' perceptions of test requirements and preparation activities are not always in line with those of their trainers (Weir and Green, 2002). Test takers may use strategies that they do not usually use in non-testing situations, and these strategies may not be what the test purports to measure (Cohen & Upton, 2007). This study looked into different perspectives from test takers, test trainers, and test developers with a view to uncovering their opinions of taking, training and developing an English test. The aspects examined include the general test (task) difficulty, test-taking strategies, the effects of test preparation, as well as the impression of the particular test under investigation. The primary interest of this study is to disclose discrepancies (if any) between the opinions and practices of the three lines of test stakeholders in order to further build the validity arguments of the test. Five test-takers, three test trainers, and one test developer were interviewed. The findings of the study revealed that in successfully answering test questions, test-takers were taught by trainers to use some strategies which might not be in congruence with what test developers intended to assess. In addition, there were some general test-taking strategies

which test takers were instructed to use, but it was up to the students to use the strategies they feel comfortable with in the testing situation. No fixed cognitive processing procedures could be easily identified that could work for everyone. This was especially true in the case of answering integrative item types, which were considered difficult but good tasks as test-takers and trainers believed such tasks assess real language ability.

**15:00-15:30 Paper 9
Yang Lu**

**Exploring the criterion-related
validity of New HSK for a U.K.
university's Mandarin Chinese
courses**

HSK is a standardised proficiency test for Mandarin Chinese which has now more than 100,000 candidates each year. The New HSK was launched in 2009 and has claimed to strive for “test-teaching correlation” (2011, chinesetesting.org.cn). Mandarin Chinese has become one of the modern languages that more and more students in U.K. universities are studying as part of their degrees. As a result, the New HSK has been the standardised and most widely acknowledged proficiency test for Mandarin Chinese that students take to certify their levels in the language for professional and academic ambitions. The situation has posed a mission for the Mandarin Chinese language programmes in U.K. HE to relate their assessments to the criteria and levels of proficiency by the New HSK. This paper, therefore, attempts to investigate the criterion-validity of the New HSK and explore the agreement between the performance of the students on the institutional assessment and that on the New HSK exams of the same level. Descriptors of the New HSK tests at Level 3 and 4 papers are presented in comparisons with the teaching outcomes and marking criteria for the Level 3 and 4 Mandarin courses in the University of Nottingham. Scores awarded by the New HSK listening, reading and writing papers to the subjects and the marks given to them in the achievement tests of the language programme are examined to see the degree that the two measures correlate and the New HSK's criterion-related validity in U.K. HE settings. Findings suggest that greater discrepancy between the criteria by the New HSK and those by the university at Level 3 and higher correlation seem to exist at Level 4. The paper will discuss the diplomas in testing Mandarin Chinese which may have caused the discrepancy and problems for the New HSK's criterion-related validity.

15:30-16:00

Tea/Coffee Break

16:00-16:30	Paper 10 Abdul Halim Abdul Raof, Anie Attan, Noor Abidah Mohd. Omar, Masputeriah Hamzah, Azian Ab. Aziz	Assessing graduating students' communication skills: Issues and challenges
--------------------	--	---

In an increasingly competitive and challenging global environment, workplace expectations are becoming more demanding. Would-be employers are looking for more than mere academic qualifications in graduates. One attribute which has been recognised to be an important facet at the workplace is the ability to communicate effectively. Tertiary institutions and training organizations have been put to task to produce trained prospective employees to meet the nation's ever increasing demand for a skilled and knowledgeable workforce. This paper reports on our attempt at developing a workplace readiness test of communication skills aimed at measuring graduating students' ability to communicate in contexts in which they would be employed. Bachman's (1990) framework of communicative language ability, emphasizing on linguistic, discoursal, pragmatic and strategic competencies became our basis for the design of the testing instruments. In this paper we shall outline some of the major issues and challenges we faced at the planning, designing and operational stages. Comments and suggestions shared during this forum should enlighten us on some of the findings and lessons learnt of similar testing exercises in other contexts. This we believe could assist us towards refining the construct of communicative language ability which we had initially conceptualized and the instruments designed to measure it.

16:30-17:00	Paper 11 Vanda Papafilippou	Discipline, punish and authority: the power of English language tests within the Greek context
--------------------	--	---

The English language testing industry has grown deep roots in Greece, roots that spread and eventually got entangled with the Greek mentality, thus rendering English language tests and their resulting certificates the main aim of the whole learning process (Papafilippou, 2012). But why have these tests become so powerful within the Greek context? This paper aims at exploring the nature of power that English language tests exert in Greece, how this power is (re)produced, and invite the audience to start questioning the values behind and the ideology delivered through English language tests (Shohamy, 2001). Drawing upon narrative interviews with Greek test-takers, parents, teachers and testers, it will be argued that the Greek society appears to be disciplinary (Foucault, 1977), since it creates a sort of social 'quarantine' for all those who do not abide by its demands for qualifications. English language tests appear to promote hierarchical power relations at both individual and institutional level and act as a Panopticon, as Greek test-takers seem to have learned "to judge themselves as if some external eye was constantly monitoring their performance" (Broadfoot, 1996, p. 68). The stories gathered indicated that English language tests' power primarily comes from their market value, the English language itself and the hegemonic human capital discourse with its accompanying neoliberal values (i.e., competitiveness, market ethic, entrepreneurialism). Hence, English language tests in Greece appear to exercise both symbolic (Bourdieu, 1991) and disciplinary (Foucault, 1977) power, since not only do they make Greek people to see and believe certain conceptions of the world (i.e.,

the world as globalised and competitive), but they also regulate their behaviour according to the needs of the omnipotent 'market'.

17:00-18:00	Drinks Reception (Subsidized by British Council)	4 th Floor Foyer, 35 Berkeley Square
18:00-20:30	Dinner	Jamie's, Park Street

Sunday 18 November 2012

09:00-09:30 Paper 12
John Field

**Cognitive validity in language
testing: theory and practice**

Cognitive validity in language testing is concerned with the extent to which the mental processes elicited by a test parallel those that the test taker would employ in a target communicative context. This is an especially important consideration in the case of high stakes tests which serve as predictors of performance in a professional or academic setting or of ability to reside in an L2 community. This talk will present recent developments in thinking concerning the cognitive validity of tests of L2 skills. After clarifying what is to be understood by the concept, it will go on to describe the methodological approaches that have been applied in cognitive validation studies. It will propose two generic models of the processes which underlie the language skills: one for productive skills and one for receptive. The models can be matched against information on test taker behaviour gained through verbal reports. Conclusions can then be drawn in terms of three validation criteria: How representative is a test of the cognitive processes that a test taker would employ in a real-world context? How comprehensive is the range of processes elicited by the test? How well calibrated are the processes elicited by a given test in relation to the proficiency level at which the test claims to be targeted? In considering the last question, the impact of working memory and attention upon performance will be briefly described. Examples given will apply to the testing of all skills, but there will be particular reference to the difficulties associated with the testing of second language listening.

09:30-10:00 Paper 13
Ardeshir Geranpayeh, Mark Elliot

**Validation of L2 listening
comprehension tests within a socio-
cognitive framework**

Listening comprehension is today seen as a process involving the decoding of audio input and the creation of an enriched meaning representation from the propositional content of the input interpreted within the context of the interaction (Field 2008). In this light, we will argue for the appropriacy of a socio-cognitive framework for developing and validating L2 listening tests based upon three dimensions of individual characteristics, internal cognitive processing and external contextual factors (Weir 2005), which reflects the nature of language as relating to a set of internal cognitive processes but taking place within an external social context. We will argue that a socio-cognitive framework provides a systematic and practical basis upon which to develop and validate tests of L2 listening, particularly in terms of their cognitive and contextual factors, and that conducting such validation exercises also allow us to scrutinise and refine the socio-cognitive framework in terms of its applicability to L2 listening tests. In so doing, we will make reference to the forthcoming *Studies in Language Testing* volume *Examining Listening* (Geranpayeh and Taylor forthcoming), which investigates both the theoretical issues surrounding the application of the socio-cognitive framework and its specific application to the validation of

the Cambridge ESOL Main Suite listening tests, discussing key findings and issues arising from the validation exercise. This presentation contains material of relevance to both practitioners and theorists, bringing together issues of applied linguistics and language testing, and considering the interplay between cognitive psychology, SLA and sociolinguistics.

10:00-10:30 Paper 14

**Fumiyo Nakatsuhara,
John Field**

**Examiner interventions in oral
interview tests: what are the
listening demands they make upon
candidates?**

Examiners intervene in second language oral interviews in order to elicit intended language functions, to probe a candidate's proficiency level or to keep the interaction going. Interventions of this kind can indirectly affect the candidate's performance and test score, since the candidate is obliged to process them as a listener and respond to them as a speaker. The present study examined forty audio-recorded interviews of the oral test of a major European examination board, to offer a better understanding of the listening demands that examiner interventions make upon candidates. Half of the interviews involved candidates who scored highly on the test while the other half featured low-scoring ones. This enabled a comparison of the language and behaviour of the same interviewer across proficiency levels, to see how they were modified in response to the communicative competence of the candidate. The recordings were transcribed and analysed with regard to types of examiner intervention, in terms of a) lexical and syntactic complexity, b) information density, c) the length and frequency, d) articulation rate and e) purpose of interventions. The selection of these language properties was based on Weir's (2005) socio-cognitive framework for validating listening tests. Further analysis was then carried out to examine the extent to which the interventions varied in response to the proficiency level of the candidate. The study affords a new insight into examiner-examinee interactions. It also has a concrete outcome in a set of criteria which could extend test specifications on the role of interviewers and contribute to the future training of examiners.

10:30-11:15

Tea/Coffee Break,
Group Photo (Brandon Hill, weather permitting)

11:15-11:45 Paper 15
Talia Isaacs

**Interactional patterns and
interlocutors' perspectives on a
paired speaking test task**

This study examines interactional patterns between English language learners from different first language backgrounds on a collaborative speaking task from the Business English Certificates, a standardized test often used to screen prospective employees for their English language proficiency at international companies. Peer interactions of 84 international engineering graduate students, categorized using Galaczi's (2008) interactional typology, were examined in relation to individual interlocutors' target language use in daily

life, oral proficiency level, and perceptions of their performance on the task. Results support evidence from the language assessment literature that asymmetric interactions tend to occur when there is a mismatch in paired test-takers' L2 proficiency level (May, 2009). This echoes indications from SLA classroom interaction research that learner pairs with largely discrepant L2 proficiency levels are less likely to be collaboratively oriented in their discourse productions than when they are more evenly matched (Kowal & Swain, 1994). Overall, collaborative pairs were the most satisfied with the quality of their interaction, experienced the fewest pronunciation and communication problems, and were most receptive to the idea of participating in future professional collaborations with their partner. Conversely, dominant members of unevenly matched (asymmetric) pairs were relatively less positive about the interactional outcome and their partner, likely due to poorer peer engagement (e.g., as signaled through a lack of topic extension). Implications for peer pairing practices in assessment contexts and in real world workplace and academic settings will be discussed.

11:45-12:15 Paper 16

**Gad Lim, Evelina Galaczi,
Nahal Khabbazzbashi**

**Descriptor salience and clarity in
rating scale development and
evaluation**

Rating scales in performance assessment need to meet the competing requirements of construct coverage (i.e. be relatively long and detailed) and examiner usability (i.e. be relatively short and succinct), thus making the question of what to include/exclude in scale descriptors a crucial consideration. As raters make myriad observations of performances, and descriptors are simplified articulations of those observations (Lumley, 2005), the inclusion/exclusion decision can be informed by the salience and clarity of criteria to raters. The context of this paper is a larger project reviewing the analytic speaking and writing scales of an international test of English language proficiency. A sample of examiners (Speaking n=1141; Writing n=621) completed a survey about the salience and clarity of concepts and descriptors in the rating scales. Forced response items were tabulated, open-ended items were analysed using an emergent coding approach to discover major themes, and results were triangulated. This paper reports findings from the survey for each sub-scale for each skill, addressing the following questions: Which concepts and descriptors did examiners find salient/not salient? Which concepts and descriptors did examiners find clear/not clear? Are there descriptors that are unclear vis-à-vis descriptors at adjacent score points? For criteria found in both scales (e.g. lexis, grammar), to what extent are examiner responses comparable and why? What performance features not captured in the scales did examiners think should be included? The implications of examiners' responses will be discussed in terms of construct coverage, rating reliability and rater training, and key themes which emerged for each sub-scale will be explored in more detail (e.g., the role of accent in the 'Pronunciation' sub-scale; approaches to dealing with unusual performances). The paper concludes with a discussion of the implications for the scale review project in particular and on rating scale development theory and practice in general.

12:15-13:15	Closing Talk Professor Barry O’Sullivan British Council/Roehampton University	Abstract (on page 5)
13:15-13:25	Farewell and LTF 2013	
13:25-14:25	Lunch (Subsidized by Pearson)	4 th Floor Foyer, 35 Berkeley Square

ABSTRACTS OF POSTER PRESENTATIONS

The relationship between test-takers' first language, listening proficiency and their performance on paired speaking tests

Suwimol Jaiyote University of Bedfordshire

This presentation describes a small-scale pilot study which examined the relationship between test-takers' L1, listening proficiency and their performance on paired speaking tests. 12 participants from three different L1 backgrounds participated in the study. They took two paired speaking tests: once with a shared L1 partner, and once with a non-shared L1 partner, as well as a listening test and a monologue speaking test to measure their listening ability and individual speaking ability. After each paired speaking test, the participants were also interviewed about their test taking experience. All speaking tests and interviews were video-recorded and transcribed. The results indicated that test-takers with higher listening proficiency tended to get higher scores in discourse management and interactive communication categories in paired speaking tests. They tended to get higher pronunciation scores when paired with non-shared L1 partners. Analysis of the speaking test discourse and retrospective interviews elaborated on these statistical results.

Teachers' perspective of washback effects of English reading tests in Japanese university entrance examinations

Yumiko Moore University of Exeter

Language tests, especially high-stakes tests, have an impact on teaching and learning, so-called "washback effect of language tests". In general, language tests are criticised for having negative impact in the way that students in language classes are practicing examination techniques rather than language learning activities. However, some research on washback effect of tests find that language teaching methodology is affected by not only examinations by but also various factors: teachers' beliefs and preference, pressure from students, their parents and principal in schools. In order to find whether Japanese university entrance examinations have a negative washback on teaching, this study examines the content and construct validity of reading comprehension tests in Japanese university entrance examinations in terms of the appropriateness of examinations for testing English language skill, particularly reading skills. This study also examines the consistency between reading skills measured in reading tests in university entrance examinations and those taught in a reading class which are mentioned in the National Curriculum, the Course of Study. This study finally suggests how Japanese teachers of English perceive washback effect of examinations on their teaching methodology through a semi-structured questionnaire.

An investigation into the construct validity of an academic writing test in English with special reference to the Academic Writing module of the IELTS test

Ahmad Alsagoafi University of Exeter

The International English Language Testing System (IELTS) is the world's leading high stakes test that assesses the English Language Proficiency of candidates who speak languages other than English and wish to gain entry into universities where English is the language of instruction. Acknowledging this popularity and importance to learners of English as a Foreign Language (EFL), this poster gives a general overview of this qualitative study which investigated the construct validity of the academic writing module in the IELTS test from the eyes of the stakeholders. The aim was to understand why Saudi students fail to cope with demands of the university despite the fact that they have achieved the minimum requirements in IELTS. Data was collected in two phases in two different settings through questionnaires, observations and interviews. The sample of the study included: 8 students, 6 university lecturers and markers. Preliminary results have shown that there was a low level of correspondence between IELTS tasks and university tasks in terms of level of difficulty, variety of topics, styles, discourse mode, length, and language. Moreover, stakeholders were doubtful about the issue of readiness of students, which is claimed by IELTS, and they wanted the test to be clearer about how the students are going to cope with university demands upon gaining entry. Furthermore, this study found that the misunderstanding of the requirements of the task could also be another important factor affecting the students' performance.

Korean teachers' attitudes toward the National English Ability Test in Korea: Implications for washback

Kyeonghwa Lee University of Bristol

The consequences of testing, inter alia, high-stakes exam on not only teaching and learning but also the education system and society have received more and more attention. Widely known as washback, a great deal of research has been conducted to examine the testing impact on teaching and learning such as changes in teaching methods, content and students' performance. However, little is done on educational and social phenomenon. As a high-stake exam in Korea, College Scholastic Aptitude Test (CSAT) assessing only reading and listening has come under criticism for students' lack of communicative skills, a narrowing curriculum, and encouraging cramming or private tutoring. In response to such issues the National English Ability Test (NEAT) is introduced to promote positive washback effects and regulating the increasing private education. The present research focuses on investigating the viewpoint of teachers in Korea with particular reference to the impact of NEAT on teaching and learning, governmental policy to regulate the growth of private tutoring and the fundamental support for successful implementation. The findings suggest that NEAT induces a noticeable increase in emphasis on speaking and writing skills. In discussing teachers' attitudes towards NEAT in relation to private education, most of the teachers expressed negative opinions about NEAT as a means of regulating the phenomenon. Lastly, although useful, the teacher training should be more practical in reference to teaching methods applicable to the classroom. Furthermore, due to the shift from paper based to internet based, facilities such as computers and teaching materials are considered the most essential support for now. There is a need for further research into different stakeholders

involving students, parents and policy makers. Moreover, it is recommended that a study based on classroom observation be conducted to explore the extent how teaching and learning is improved by NEAT.

Suitability of Pearson PTE Academic as an exit test for a course of academic English

Nathaniel Owen University of Leicester

This study addresses the question of whether or not the Pearson Test of Academic English (PTE) is suitable as an exit test for a course of academic English language instruction, such as a university pre-sessional. The principal research question was broken down into three subsidiary questions: 1.) Are decisions made on the basis of participants' test scores in the Pearson PTE Academic and the ELTU exit test comparable? 2.) Are the contents of the two tests comparable? 3.) Are the facets of the two tests comparable? The principal methodology involved 55 members of the English Language teaching Unit (ELTU) taking the Pearson PTE examination, funding for which was provided by Pearson Education Ltd. The first research question was answered by directly comparing decisions made on the basis of performance on both tests. The second research question considered the construct of 'academic English' as determined by both sets of test writers through an analysis of test content. Test content was analysed through a process of reverse engineering (RE), realised via a principled approach of Robert Mislevy's evidence-centred-design (ECD) and Bachman's test method facets (TMFs). This was supplemented by the third research question: three expert practitioners (two of whom were tutors at the ELTU, the other was a member of the School of Education at Leicester University) were asked to rate items across both tests in relation to a taxonomy of language skills as devised by Bachman (1990). Means and standard deviations were calculated from these ratings. Where the differences between raters' means exceeded the standard deviations, this suggested that there significant disagreement between those raters regarding the skills that these items assessed.

The English test effect of the Thai university entrance examination on English teachers and students

Chamnan Para University of Bristol

In Thailand, there is the national university examination, O-NET (the Ordinary National Educational Test) launched in 2008 to replace the former examination. The scores from O-NET are one of the required scores in addition to other standardized tests and students' school grades. The English test in O-NET has long been widely criticized for its impact. This study aimed to investigate the washback of the English test in the O-NET to unfold the washback to both policy and practice levels. The data collection, employing the mixed methods, was conducted in October in 2011 to February in 2012. It was to collect the empirical data from 'classroom observations', 'semi-structure interviews', 'focus groups' and 'questionnaires'. In this study, there are two major groups of participants: four English teachers and 38 students (grade 12 students) in two Thai high schools, School A and B, in a province of Thailand. The preliminary results of this study indicated that the majority of participants were negatively affected. For two English teachers in School A, their teaching became more test oriented. Their English materials and lessons were designed to gear toward the test items in the English test of the O-NET. Accordingly, they provided students with all possible test-relating contents and necessary test-taking strategies to boost the

English test scores. In contrast, the two teachers in school B, they still strictly taught English to students in line with the assigned curriculum. In their English classrooms, all four English skills were proportionately provided and balanced. They had a firm belief in their teaching quality that it was sufficient for students to use English for all purposes, including taking the English test in O-NET. Regarding the washback to students, students in both schools shared the similarity. They tended to be more selective in the English content they study. Students intentionally drilled the content that was likely to appear in the test and disregarded the unlikely ones. For example, they recited specific grammar points and vocabulary appeared in the previous English tests. They dropped some speaking and writing skills because they were not tested. In some cases, students skipped normal classes to attend private tutorial schools in order to have a full test preparation.

Automated assessment of lexical diversity and Ngrams in essays at different levels of the CEFR Jeanine Treffers-Daller, Shirley Williams, Patrick Parslow, University of Reading

The current project contributes to the ongoing discussion (Hulstijn, Alderson and Schoonen 2010) about the identification of criterial features in learners' written output which discriminate between the levels of the Common European Framework of Reference for Languages (CEFR). We focus on how automated analyses of vocabulary usage in essays of EAP learners can further our understanding of students' achievement at different levels of the Pearson Test of English Academic (PTE Academic), which is based on the CEFR. Analyses revealed significant differences between lexical diversity scores D (Malvern et al 2004) and HDD (McCarthy and Jarvis 2007) calculated on the written work of 178 students whose essays corresponded to the standards for B1, B2, C1 and C2 ($F(3,174) = 6.92, p < .001$), but these measures were not able to discriminate between all the CEFR levels. Analyses of students' use of ngrams, that is word groups of different word lengths (Hussey et al. 2011, 2012), showed that ngrams up to 20 words were repeated in the data, so we plotted the total number of ngrams in the corpus for each value of n (1 to 20) against the number of ngrams. This revealed two 'linear' elements with a clear turning point between them, namely the 6gram point, which is important as many studies seem to limit investigations to 4grams (e.g. Yannakoudakis et al 2012). Preliminary analyses of the relationship between ngrams of up to ten words and the CEFR levels revealed that there is a slight increase in their use from B1 until C2. We will also present analyses of the differences between vocabulary deployment in the learner data and the Pearson International Corpus of Academic English (PICA-E, Ackermann, De Jong, Kilgariff and Tugwell (2010)).

Medical student's perceptions of fairness in written exams

Enas Elmansuri University of Bristol

English has been used as the medium of instruction in all the higher education scientific institutions in Libya such as the Faculty of Medicine in Tripoli. Medical students have to sit for examinations written in English. Written exams in English not only require the students' medical knowledge but also their English language proficiency. This raises a series of issues such as:

- to what extent do medical students feel the examinations are fair,
- what is the students' knowledge or understanding of assessment of fairness,
- what are their expectations of a fair examination,
- what do students base their fairness judgments on, and
- what could be done to improve fairness of these written exams?

This PhD research is still in progress. This research investigates the medical students' perceptions of fairness in a written surgery examination in Libya. It draws on fairness research in both general educational assessment and language assessment. This research benefits from using mixed method for data collection, including a measure of fairness through a questionnaire followed by semi-structured interviews with participants and critical analysis of documents and literature. The participants in the study are the fifth year medical students who are taking surgery subject at Tripoli University. The findings of this research will help better understand (a) the role of English language in assessing students' subject knowledge, (b) the role of language in understanding test fairness in educational assessment for specific purposes, and (c) the support for English language teaching in medical schools in Libya and similar contexts in other education systems.

A Thinkpiece - ESL curriculum and testing for vocational purposes: Is it time to bring it all together?

Maria Doyle

SELECT Consultants, Australia

Following is a think-piece which is headlining a national movement in Australia to standardize the widely used ISLPR language testing system. More detailed information on the ISLPR can be found at: <http://www.islpr.org>.

ISLPR (International Second Language Proficiency Rating) and IELTS have been the two major testing systems used in Australia for the past thirty years; IELTS since 1989, and ISLPR since 1979. The nature of ISLPR and IELTS vary considerably although they both seek to rate a candidate's performance on the four macro skills of speaking, listening, reading and writing. Wylie (2007) describes the difference in origins as:

*'Development of the **ISLPR** began in 1978. The first version of the scale (originally called ASLPR) was developed by a Queensland Working Party led by David Ingram and Elaine Wylie, with a Steering Committee composed of representatives of Commonwealth (Immigration and Education) and State (Education) departments ... Development of **IELTS** began in 1987. The development team was based at the University of Lancaster. It consisted of three applied linguists, two British – Charles Alderson and Caroline Clapham – and one Australian – David Ingram. There was a Steering Committee composed of representatives from the UK and Australia.'*

For the purposes of whether IELTS or ISLPR is more appropriate for the TVET (Technical and Vocational Education and Training) environment, differing factors such as test development, authenticity of tasks, and test administration must be taken into consideration. In terms of test development;

'ISLPR testing is adaptive, or custom built, to the individual candidate. Speaking, listening and reading are judged in a one-to-one interview. With writing, a number

of candidates may do the test in the same room at the same time, but generally they will have different tasks. There is a high level of standardisation in the IELTS testing process. Listening, reading and writing are common tests, administered en masse. Speaking is tested one-to-one and the tester chooses from a range of tasks provided.'(Wylie, 2007)

IELTS has enjoyed success as the leading test accepted by DIAC (Department of Immigration and Citizenship, Australia), tertiary institutions and employers Australia- wide and internationally, for immigration and employment purposes.

ISLPR has been accepted by a number of industry bodies around Australia and is the system on which the CSWE (Certificates in Spoken and Written English), an AQF (Australian Qualifications Framework) compliant curriculum developed by NSW AMES is based. The courses are also used widely throughout Australia in the Adult and Community Education (ACE) sector, TAFE (Tertiary and Further Education) colleges, university language programs, Indigenous education, corrective services programs, private language colleges and Intensive English Centres (IECs) in NSW state schools. The CSWEs are also widely used by service providers of the AMEP (Adult Migrant Education Program), which is a migrant settlement program provided by the Australian Government and administered through the Department of Immigration and Citizenship. The NCELTR (National Centre for English Language Teaching and Research, based at Macquarie University) also produced an online, moderated assessment task bank, which was created with the help of staff from numerous institutions delivering the CSWE Australia-wide. TAFE Institutions Australia wide also use ISLPR to grade their students into levels and train their English teachers in using the ISLPR to grade, monitor and assess learners progressing through CSWE certificates.

When comparing the systems, it is apparent that, although both systems are highly regarded and have been subject to thirty years of evaluation and development, there are cohorts of candidates better suited to each system. In terms of the adaptability and personalization of test development to suit the context of the candidate, it would seem that ISLPR is the testing system better suited to cohorts of refugees, migrant workers and immigrants; ISLPR also has the ability to rate the candidate based on their ability to communicate in certain vocational, academic or general contexts.

Which poses the question: if ISLPR has, for the last thirty years, been the basis for the in-house testing and teaching of a range of candidates from academic, non-academic, migrant, refugee, indigenous and vocational backgrounds, is it not time to formally register ISLPR as a nationally accredited testing system, governed by an independent body which can moderate and validate testing and resulting? Is it not time for the nation to formally recognize and accept ISLPR as an acceptable alternative for institutional or immigration requirements? Is it not time to make language testing fairer for non-academic and vocational candidates, so that they, too, can access testing which can indicate their true ability to communicate in their chosen field?

Research questions arising from this movement might include the validity and reliability of major testing systems for minority groups including but not limited to refugees, migrants, indigenous candidates and non-academic or vocational candidates.

CAER

**Centre for Assessment and
Evaluation Research**

Graduate School of Education
35 Berkeley Square,
Bristol, BS81JA UK

bristol.ac.uk/education