<table>
<thead>
<tr>
<th>Day</th>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tuesday - Thursday, May 21st - 23rd</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Pre-conference Workshops</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Workshop 1</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Workshop 2</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Workshop 3</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Sakarya A</strong></td>
</tr>
<tr>
<td>Tues.14.00 -</td>
<td></td>
<td>The Impact of Academic Writing Tests</td>
</tr>
<tr>
<td>Thurs.12.15</td>
<td></td>
<td><em>Tony Green, University of Bedfordshire, UK</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Sakarya C</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Testing Vocabulary</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Diane Schmitt, Nottingham Trent University, UK</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Sakarya D</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Classroom-Based Language Assessment</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Neus Figueres Casanovas, Oscar Soler Canela, Dina Tsagari</em></td>
</tr>
<tr>
<td></td>
<td>12.30-13.30</td>
<td>Early registration</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>SIG Meetings</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Barbaros A</strong></td>
</tr>
<tr>
<td></td>
<td>13.30-17.30</td>
<td>Classroom-Based Language Assessment</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dina Tsagari</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Barbaros B</strong></td>
</tr>
<tr>
<td></td>
<td>13.30-17.30</td>
<td>Assessing Writing and Assessing for Academic Purposes (joint meeting)</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Claudia Harsch, Peter Holt &amp; Filiz Etiz</em></td>
</tr>
<tr>
<td></td>
<td>17.30-19.00</td>
<td>Registration</td>
</tr>
<tr>
<td></td>
<td>18.30-20.30</td>
<td>Social event: <em>Opening Reception at the Istanbul Military Museum, Harbiye, Istanbul</em> Sponsored by Pearson</td>
</tr>
<tr>
<td><strong>Friday, May 24th</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.15-8.50</td>
<td>Registration</td>
</tr>
<tr>
<td></td>
<td>8.50-9.00</td>
<td>Opening remarks</td>
</tr>
<tr>
<td></td>
<td>9.00-10.00</td>
<td>Plenary 1: Professor Liz-Hamp Lyons</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>The challenge of classroom-based assessment</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Papers (Main Hall)</strong></td>
</tr>
<tr>
<td></td>
<td>10.00-10.30</td>
<td>The impact of TOEPAS as a language policy tool</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Slobodanka Dimova</em></td>
</tr>
<tr>
<td></td>
<td>10.30-11.00</td>
<td>The consequences of diagnostic assessment on teaching and learning</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Roxanne Wong &amp; Michelle Reyes Raquel</em></td>
</tr>
<tr>
<td></td>
<td>11.00-11.30</td>
<td>Coffee Break</td>
</tr>
<tr>
<td></td>
<td>11.30-12.00</td>
<td>What are we aligning tests to when we report test-alignment to the CEFR?</td>
</tr>
<tr>
<td></td>
<td>12.00-12.30</td>
<td>The impact of institutional tests on teachers’ choice of instructional strategies</td>
</tr>
<tr>
<td></td>
<td>12.30-13.00</td>
<td>Assessing students with disabilities: voices from the stakeholder community</td>
</tr>
<tr>
<td></td>
<td>13.00-14.30</td>
<td>Lunch and Poster Session (see page 25-27 for details)</td>
</tr>
</tbody>
</table>
## Parallel Paper Presentations

<table>
<thead>
<tr>
<th>Time</th>
<th>Main Hall</th>
<th>Barbaros A</th>
<th>Barbaros B</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.30-14.55</td>
<td>Evaluating the impact of applying indigenous assessment criteria in an LSP test <em>Cathie Elder</em></td>
<td>Understanding the role of subject-matter knowledge in ESP ability: messages for teachers <em>Yuyang Cai</em></td>
<td>The impact of a test preparation program as an embedded component of the syllabus <em>Fatma Şentürk</em></td>
</tr>
<tr>
<td>15.00-15.25</td>
<td>School teachers and trained raters assessing L2 writing of the same students <em>Mirja Tarnanen</em></td>
<td>Teacher positioning towards high-stakes exams in the ESOL sector <em>Simon Webster</em></td>
<td>TEMPUS Proset: The impact of a language assessment course in a micro-educational context <em>Marina Solnyshkina</em></td>
</tr>
<tr>
<td>15.30-15.55</td>
<td>Acting under the guidance of subject-matter knowledge in ESP ability: messages for teachers <em>Yuyang Cai</em></td>
<td>A task-based approach to investigating the language proficiency needed to teach EFL</td>
<td>Torn in two directions? Functional failure in dual function high stakes exams <em>Tania Horak</em></td>
</tr>
</tbody>
</table>

### Coffee Break

**Parallel Work-in-Progress Presentations**

<table>
<thead>
<tr>
<th>Time</th>
<th>Main Hall</th>
<th>Barbaros A</th>
<th>Barbaros B</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.30-16.50</td>
<td>The impact of using graphic organizers to test reading comprehension <em>Marci Nelson Özer</em></td>
<td>Monitoring the effects of exam renewal: teachers views on preparation material <em>Jayanti Banerjee &amp; Barbara Dame</em></td>
<td>The impact of standardised testing on the teaching of English to teenagers <em>Claudia Mewald</em></td>
</tr>
<tr>
<td>16.55-17.15</td>
<td>Game Over? – Investigating the impact of ECL exams on successful candidates <em>Róbert Márcz &amp; Gábor Szabó</em></td>
<td>Examining washback in multi-exam preparation classes in Greece <em>Irini Papakammenou</em></td>
<td>Listening tests-the impact of mode of delivery on performance <em>Eli Moe</em></td>
</tr>
<tr>
<td>17.20-17.40</td>
<td>Making the most of the CEFR: developing an assessment programme beyond Europe <em>Kathryn Brennan</em></td>
<td>Computerized ToEFL exam acceptance <em>Kubra Karakaya Özyer &amp; Beyza Dunya</em></td>
<td></td>
</tr>
<tr>
<td>18.30-22.30</td>
<td>Social Event: <em>Evening on the Grande Hotel de Londres Terrace</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Saturday, May 25th

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 9.00-10.00 | Plenary 2: Professor Art Graesser  
*The impact of automated measurement of text characteristics* |
| 10.00-10.30| A national exam’s washback on reading assessment in the secondary classroom  
*Doris Fröestscher* |
| 10.30-11.00| Getting ready: a study of the impact of test preparation  
*Glyn Jones* |
| 11.00-11.30| Coffee Break |

#### Papers (Main Hall)

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
</tr>
</thead>
</table>
| 11.30-11.55| The washback of new high-stake examinations in English in Poland  
*Elżbieta Zawadowska* |
| 12.00-12.25| Developing and validating band levels for reporting overall examinee performance  
*Xiaoming Xi* |
| 12.30-12.55| Exploring the criterion-related validity of the new HSK and its impact  
*Yang Lu* |
| 12.55-14.30| Lunch and Poster Session (see page 25-27 for details) |

#### Parallel Paper Presentations

<table>
<thead>
<tr>
<th>Main Hall</th>
<th>Barbaros A</th>
<th>Barbaros B</th>
</tr>
</thead>
</table>
| 11.30-11.55                | The impact of classroom assessment on Thai EFL undergraduate students  
*Jirada Wudthayagorn* | Item development, the CEFR, and the perils of Cinderella testing  
*Thom Kiddle* |
| 12.00-12.25                | Examining the consequential validity of a standardised test in a classroom setting  
*Christine Doe* | Fairness in the secondary schools: A Singapore case  
*Chih-Min Shih & Li-Yi Wang* |
| 12.30-12.55                | Using and interpreting English proficiency tests scores in the graduate admission process  
*April Ginther* | Peer evaluation impact on testing oral presentation skills  
*Okan Önalan* |

#### Papers (Main Hall)

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
</tr>
</thead>
</table>
| 14.30-15.00| Demonstrating validity for high-impact language tests  
*Nivja de Jong & Jelle Goeman* |
| 15.00-15.30| An interpretive argument for automated writing evaluation as a formative assessment tool  
*Stephanie Link & Ahmet Dursun* |
| 15.30-16.00| Coffee Break |
| 16.00-17.30| Annual General Meeting  
(Main Hall) |
| 19.00-23.30| Social event: Bosphorus cruise and conference dinner |
Sunday, May 26th

### Papers (Main Hall)

<table>
<thead>
<tr>
<th>Time</th>
<th>Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.30-10.00</td>
<td>What is the impact of diagnostic language tests? Ari Huhta</td>
</tr>
<tr>
<td>10.00-10.30</td>
<td>Task design for automatic scoring of interactive spoken pragmatic skill Jared Bernstein</td>
</tr>
<tr>
<td>10.30-11.00</td>
<td>Factors affecting long-term examination impact and the fate of the examinations themselves Dianne Wall</td>
</tr>
<tr>
<td>11.00 - 11.30</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11.30-12.30</td>
<td>Plenary 3: Professor Dan Douglas Nobody Speaks English Here Today: The Notion of Impact in the Context of Aviation English</td>
</tr>
</tbody>
</table>

### Conference Close

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.30-18.00</td>
<td>Social event: Istanbul city tour</td>
</tr>
</tbody>
</table>

### Poster Sessions

**(Friday and Saturday, 13.00-14.30, Main Foyer)**

<table>
<thead>
<tr>
<th>Poster Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flipped washback? Improvisational theatre as a means of preparing dialogical oral assessment</td>
</tr>
<tr>
<td><em>Karin Vogt &amp; Raphaelle Beecroft</em></td>
</tr>
<tr>
<td>Tests for Turkish as a heritage language</td>
</tr>
<tr>
<td><em>Sibylle Plassmann</em></td>
</tr>
<tr>
<td>Righting writing practices: The impact of exam reform</td>
</tr>
<tr>
<td><em>Kremmel Benjamin, Kathrin Eberharter &amp; Eva Konrad</em></td>
</tr>
<tr>
<td>The impact of gender on interviewer behaviour during oral proficiency interviews</td>
</tr>
<tr>
<td><em>EneAlas &amp; SulikoLit</em></td>
</tr>
<tr>
<td>Comparing two modes of test administration</td>
</tr>
<tr>
<td><em>Linda Nepivodová</em></td>
</tr>
<tr>
<td>The English test washback of Thai university entrance examination on teachers and students</td>
</tr>
<tr>
<td><em>Chamnan Para</em></td>
</tr>
<tr>
<td>Exploring the Washback Effects of High-Stakes Tests on ELT in Russia.</td>
</tr>
<tr>
<td><em>Ludmila Kozhevnikova &amp; Viktoria Levchenko</em></td>
</tr>
<tr>
<td>The impact of the speaking component of an institutional test on students’ anxiety level</td>
</tr>
<tr>
<td><em>Selen Sökmen</em></td>
</tr>
</tbody>
</table>
The Challenge of Classroom-Based Assessment

Stiggins (2001) tells “a story of disregard by policy makers, school leaders, and the measurement community” for classroom assessment that has “kept classroom assessment from even approximating its potential” (p.5) But in the decade-plus since he wrote those words, a good deal of attention in the field of educational assessment has been turned to various forms of assessment that are not ‘testing’: formative assessment; assessment for learning; learning-oriented assessment; school-based assessment; classroom-based assessment (and more). The first challenge then, in talking about classroom-based assessment (CBA), is to untangle the terminology in order to get to the heart of the question: what is classroom-based assessment?

The second and more probably more relevant challenge will be to formulate at least some principles that are essential for ‘classroom-based assessment’ to be qualitatively and effectively different from traditional testing. For example: Is classroom-based assessment necessarily/always formative? Is overt, explicit feedback a necessary element of classroom-based assessment? Can assessment be “for learning” (as in Black & Wiliam 1998; Black et al 2003) if it is taking place in the classroom but being reported upwards to accountability systems?

The third challenge in the context of a group such as EALTA is to consider whether (and perhaps which of) the principles hold true when working with second, third or fourth language learners, and also whether there may be some additional principles that apply specifically to the teaching and learning of languages. The fourth challenge is perhaps the greatest: since classroom-based assessment is at once new and different, yet at the heart of what good teachers already do, have in fact always done, how can teachers learn and practice the skills of assessment in their classrooms? I will argue that greater attention must be paid to teacher, student and administrator assessment literacy if relatively sophisticated alternative assessment tools and strategies are to be implemented successfully, and if the changes implicit and explicit in classroom-based assessment approaches are to have impact in students’ command of English and their attitude to language learning.


Liz Hamp-Lyons (PhD, University of Edinburgh) has worked in England, Scotland, Greece, Iran, Malaysia, the USA (Illinois, Michigan and Colorado), Australia, and in Hong Kong, where she was Chair Professor and Director of the Asian Centre for Language Assessment Research (ACLR) at Hong Kong Polytechnic University. She is a Visiting Professor in English Language Assessment at Creira, University of Bedfordshire, and a Guest Professor in International Studies at Shanghai Jiao Tong University, where she is the senior consultant to the College English Test and the Profile of Academic Competence in English. Her assessment interests are in writing and speaking assessment and in school/classroom-based assessment. Liz is also the founding Editor of the Journal of English for Academic Purposes, and Editor of Assessing Writing.
10.00-10.30  The impact of TOEPAS as a language policy tool
Slobodanka Dimova, Copenhagen University, Denmark

The increased number of English-medium programs at universities across Europe has lead to development of language policies for quality assurance of instruction. These policies are often enforced by implementation of internal procedures for assessing non-native English speaking lecturers’ oral communicative competencies. The Test of Oral English Proficiency for Academic Staff (TOEPAS), a performance-based oral English certification for university lecturers, was designed to respond to such policy at a large Nordic university. To identify the impact of TOEPAS as a language policy tool, the present study examined (1) the ways the university language policy was enforced through the test, (2) the relationship between the intended and actual test effects, and (3) the test score uses and consequences. The study was based on 30 audio-recorded interviews of randomly selected university lecturers and administrators. A content-based analysis was applied to analyze the transcribed and coded interview data (Krippendorff, 2003). Findings suggested that TOEPAS uses varied among departments and programs due to different policy interpretations. Although one of the test’s intended purposes was further language learning and development, lecturers rarely used the detailed certification reports for formative purposes. Finally, while the certification didn’t seem to affect professors in terms of their positions and teaching responsibilities, post-doctoral students and other short-term lecturers showed some anxiety about the consequences of not being certified. These findings will be discussed with specific reference to the micro- and macro-contexts in which TOEPAS is used (Watanabe, 2004).

10.30-11.00  The consequences of diagnostic assessment on teaching and learning
Roxanne Wong, City University of Hong Kong
Michelle Reyes Raquel, The Hong Kong Polytechnic University

This paper identifies the impact of diagnostic assessment on teaching and learning at three tertiary institutions in Hong Kong. These institutions developed the Diagnostic English Language Tracking Assessment (DELTA) to provide learners with a tool to diagnose and track students’ English proficiency throughout their university studies. Students take the DELTA at the beginning of the academic year and receive a diagnostic e-report, which shows their ability level on a proficiency scale. It also shows their strengths and weaknesses in reading, listening, vocabulary, grammar, and writing, the skills they need to work on in order to progress, as well as links to online resources. Language instructors have access to these e-reports and DELTA teaching resources to be used as appropriate for their context.

Because of Hong Kong’s examination-oriented culture however, low student uptake of DELTA, a zero-stakes assessment, is expected. Institutions support the DELTA initiative because they realise that diagnosis of their academic English ability at the beginning of their university studies can empower students to take charge of their own learning over their four years at university. This study investigates students’ and teachers’ perceptions of the DELTA, the e-report provided, and teaching resources as tools for learning. Analysis of survey results, classroom observations, independent learning plans, student focus groups, and teacher interviews reveal the tensions involved in using DELTA for independent learning and classroom teaching. The paper concludes with an evaluation of the usefulness of a diagnostic assessment to foster second language learning in Hong Kong tertiary institutions.

11.00-11.30  COFFEE BREAK

11.30-12.00  What are we aligning tests to when we report test-alignment to the CEFR?
Claudia Harsch, University of Warwick, UK
Johannes Hartig, DIPF, Frankfurt/M, Germany

Data from a workshop aligning tests to the CEFR raised issues with the validity of the alignment, some of which have not yet been examined in enough detail. Listening tests operationalising a set of pre-defined difficulty-determining characteristics (six tasks, 82 items, characteristics regarding linguistic level of input, speed of delivery, cognitive operations and item format) were to be judged with regards to which specific CEFR-descriptors they were operationalising. We employed basket- and item-descriptor methods in two judgement rounds. Ten judges stated the CEFR-descriptors they thought most likely to be operationalised by each item, and a global CEFR-level per item. We compared judgement agreement on the CEFR-level ratings and the CEFR-descriptors quoted. We also examined the correlations between CEFR-level ratings, the pre-determined difficulty-characteristics and empirical item difficulties to shed light on which criteria the judges relied on. Follow-up interviews allowed
investigating judges’ perception of issues and problems.

Our findings show that judges relied on different characteristics and CEFR-descriptors to a varying degree. There seems little comparability in what aspects judges used to form their global CEFR-judgments. Different judges seemed to define CEFR-levels differently. Therefore, if an alignment does not take into account the meaning of the CEFR-levels as expressed by CEFR-descriptors, this raises issues with alignment validity, and hence the validity of test-score interpretations and usage. Given the impact of using CEFR-aligned tests for university admissions or immigration purposes, it is time to shed more light on what assigning a global CEFR-level to a test actually means.

12.00-12.30 The impact of institutional tests on teachers’ choice of instructional strategies

Songül Yeliz Sarı, Çağ University, Turkey

Teachers’ approaches to teaching greatly influence students’ approaches to learning, thus playing a critical role in student achievement. Based on the notion that a person’s actions are primarily determined by their beliefs, current literature emphasizes the role of teachers’ beliefs in their choice of instructional strategies. Another frequently mentioned factor that influences teacher behavior is testing. Considerable research confirms that teachers often adapt teaching materials to match the requirements of a given test. However, little has been revealed about the effects of tests on teachers’ choice of instructional strategies, mainly due to limited focus on teacher beliefs. Given the complicated nature of teacher beliefs and their far-reaching effects on teacher behavior, it is essential that they be identified in order to gain a clearer understanding of the effects of institutional tests.

The aim of this case study is to explore whether institutional testing practices influence teachers’ choice of instructional methods with special focus on teacher beliefs. In order to reach this aim, pedagogical beliefs of two native and two non-native speaker teachers of English have been identified through Rep-Grid technique, followed by classroom observations and semi-structured interviews. The classroom observations and interviews have helped us explore the relationship between teachers’ pedagogical beliefs and their actual classroom practices and whether and to what extent institutional tests have an effect on teachers’ choice of certain instructional methods. It is expected that the findings will provide some in-depth insight into the complicated relationship between testing, teacher beliefs and teachers’ choice of instructional strategies.

12.30-13.00 Assessing students with disabilities: voices from the stakeholder community

Hanan Khalifa and Lynda Taylor, University of Cambridge ESOL Examinations, UK

Growing numbers of students with disabilities are learning a second language nowadays - leading to increased demand for appropriately adapted language tests. Although language test accommodations are well established, the impact of modified assessment provision on key stakeholders in the testing process remains relatively under-researched.

This paper reports a recent empirical study to investigate issues of impact surrounding the accommodations provision of one major international examination board. The perspectives of different stakeholder groups are explored, both the test takers themselves and those directly involved in responding to a request for a modified test (e.g. exam board personnel), in preparing suitable test materials (e.g. test writers), in setting up the test venue appropriately (e.g. test centre administrators) and in conducting a modified test (e.g. speaking examiners).

Using a mixed methods approach, a review of the board’s current documentation and procedures was triangulated with survey and interview data from exam board personnel, test centre staff, test writers and speaking examiners to investigate the match between policy and practice, and to identify any issues arising. The relationship between the examination board and professional bodies/specialists with expertise in specific areas of disability was examined to understand this important partnership, especially given the growing use of assistive technology among students with disabilities. Finally, a small number of case studies were conducted with individual test takers receiving language test accommodations. The research outcomes deepen our understanding of the complex issues and challenges involved in offering appropriate and useful test accommodations for language learners with disabilities.

13.00-14.30 LUNCH AND POSTER SESSIONS
Evaluating the impact of applying indigenous assessment criteria in an LSP test
Catherine Elder, Sally O’Hagan and John Pill, University of Melbourne, Australia
Ying Zhang, Occupational English Test Centre, Australia

Criticism of specific-purpose language (LSP) tests is often directed at their limited ability to represent fully the demands of the target language use situation. Such criticisms extend to the criteria used to assess test performance, which may fail to capture what matters to participants in the domain of interest. This paper explores the implications of expanding the construct underlying assessment criteria in the context of the speaking sub-test of the Occupational English Test (OET), an LSP test for overseas-trained health professionals applying to practise in an English-speaking context.

The current OET assessment criteria are clearly linguistically oriented and applied by specially trained language professionals who are not domain experts. In order to realign the current test criteria to reflect better the requirements of effective interaction with patients, “indigenous criteria” (Jacoby & McNamara, 1999) – aspects of performance valued, in this case, by health educators in commentary on clinical interactions between trainees and patients – were identified (Elder et al., 2012).

Two of these indigenous criteria were considered amenable to inclusion in a language test: 1) management of interaction in the consultation and 2) clinician engagement with the patient. A checklist to aid understanding of these new criteria was provided for seven OET assessors at a training workshop. The assessors subsequently applied the checklist and revised criteria while rating 300 sample test performances. These ratings and assessors’ feedback following the workshop were analysed to inform further consideration of the feasibility and potential impact of using the revised assessment tool in routine test administrations.

Understanding the role of subject-matter knowledge in ESP ability: messages for teachers
Yuyang Cai, The University of Hong Kong

How teachers interpret a language test affects the influence of the test on the teaching and learning. Regrettably, dispute still exist among teachers of English for Specific Purposes (ESP) regarding the role of subject-matter knowledge. The dominating view is that there is nothing special in ESP tests but their demand on memorizing some terminologies.

Aiming to provide a better understanding of this issue, the current study examined the effect of subject-matter knowledge on ESP reading test performance by controlling terminology effect. A total of 1503 second-year nurse students from eight medical colleges across China participated by responding to the Medical and Nursing Knowledge Test measuring subject-matter knowledge and the Medical and Nursing English Reading Test measuring ESP reading ability (all medical terminologies in the texts were translated into Chinese). Data analyses involved: 1) item-level analyses; 2) bifactor multidimensional item response theory modeling (to calibrate and score the data and then to compute composite scores); 3) confirmatory factor analyses on the composite scores, and 4) full latent variable analyses to explore the predicting effect of subject-matter knowledge on ESP reading test performance. The results showed that subject-matter knowledge still had a significant effect on ESP reading test performance (b=.20, p<.000) after controlling possible terminology effect. The implication for bringing about positive washback effect was that ESP teachers should consider seriously subject-matter knowledge and assume the responsibility of helping, if not teaching, ESP learners to activate the subject-matter knowledge they bring to the ESP context.

The impact of a test preparation program as an embedded component of the syllabus
Fatma Şentürk, Çukurova Bilfen Schools, Turkey

This case study aims at investigating the impact of a high-stakes test, namely KET (Key English Test) by Cambridge ESOL, on a teacher’s teaching practices while preparing her students for the exam as an embedded component in the course syllabus for 7th graders at a private secondary school. Accordingly, the following research questions have been asked: (1) What is the nature and the scope of impact of the KET preparation program on classroom practices? (2) What differences can be observed between the main course of English and KET preparation programs regarding classroom practices? (3) What do the students think about getting prepared for an international exam? (4) What do the teachers think about preparing students for an international exam?

Recordings of a teacher’s main course of English and KET preparation classes for a period of two months have revealed the impact of getting prepared for a test on classroom practices. The recorded data
have been content analyzed to identify and compare the classroom interactions and practices in those classes, and the analysis of data have presented some similarities and differences between these classes having different foci. Interviews, conducted to reveal the two parties’ thoughts about exam preparation programs when embedded in the course syllabus, have indicated whether and to what extent they have been satisfied with this component of the syllabus.

15.00-15.25 PARALLEL PAPER PRESENTATIONS

MAIN HALL

School teachers and trained raters assessing L2 writing of the same students
Mirja Tarnanen, Ari Huhta and Tuija Hirvelä, University of Jyväskylä, Finland

In this paper, we look at how consistently trained raters and school teachers in Finland rate L2 writing of the same students of comprehensive school in grade 4-9. When collecting written performances in the schools for a second language acquisition study the second and foreign language teachers (> 300) were requested to assess their students on two different scales, one combining several Common European Framework of Reference (CEFR) writing scales, the other a national, 10-level adaptation of the CEFR scale. After data collection the performances (~2000) of the same students were rated by 25 experienced and trained raters using the two scales. Each sample was rated by 3–4 raters. The rating data were analysed with the multifaceted Rasch programme Facets. The paper will address the following questions: 1) How consistently do raters with different background (e.g. teachers vs. trained raters) rate the writing skills of the same students? 2) What do the results tell us about the familiarity of the school teachers with criterion-referenced assessment? We will present the results of the study and discuss how successfully school teachers were able to apply the relatively new form of reference assessment, which was introduced the very first time in the current National Core Curriculum for basic Education (2004). The curriculum with its new scale was intended to have a major impact on language assessment in the country but so far there has been very little research into the quality of the ratings done with the new scale.

BARBAROS A

Teacher positioning towards high-stakes exams in the ESOL sector
Simon Webster, Language Centre, University of Leeds, UK

The ESOL sector in England, which is engaged in the provision of English to adult migrants, has undergone profound changes in recent years (Hamilton 2012). As course funding becomes tied to pass rates on standardised exams, an audit culture has developed in which teachers are increasingly under pressure to achieve positive exam results (Cooke and Simpson, 2008). In addition, these exams have been adopted as a requirement for those seeking citizenship as a part of government policy. ESOL teachers therefore find themselves confronted with a high-stakes testing regime in their professional context, the impact of which has been under-researched.

This paper reports on longitudinal multiple-case study research investigating the positioning of ESOL teachers towards this exam system. The study tracks the practices and cognitions of four early career teachers over a full academic year through the use of observation and interview data. The impact of the exams on the teachers’ practical knowledge (Tsai, 2009) is explored in the context of their teaching of speaking with a strong focus on the degree to which the teachers’ practices are rationalised with reference to this feature of the institutional context. In addition, the paper discusses the range of positions adopted by the practitioners towards the exams; this includes teachers’ cognitions about the implications of the exams for the students’ immigration status and the funders’ required pass rates. Within this context, there is also an exploration of the ethical and professional issues faced by practitioners who find themselves internal verifiers for their own students.

BARBAROS B

TEMPUS Proset: The impact of a language assessment course in a micro-educational context
Marina Solnyshkina, Kazan Federal University, Russia

The report is a case study of the impact of a Language Assessment Course at Kazan Federal University piloted in 09.2012-11.2012 as part of the implementation of TEMPUS Proset (Promoting Sustainable Excellence in Testing and Assessment of English) 2011-13 Project, the major participants of which are CRELLA and 14 Russian Universities.

Main argument: LA course is a form of educational innovation and the changes on the micro-educational level need to be managed effectively. The current lack of testing and assessment components in teacher training provision in Russia cause an absence of transparent, international, educational standards in language learning and assessment in Russia. The LA course piloted at KFU is a result of the collaborative development of Consortium of 17 participants of TEMPUS Proset.

The aims of the Course are: to introduce participants to the fundamental principles of LA and develop skills in the design and validation of testing instruments.

The impact of the course is assessed at four levels: Reactions, Learning Changes, Behaviour, Results.
and include: feelings immediately after training, changes in beliefs, knowledge and skills, application of new ideas over time, effect on the organization.

**The collection of the impact data included:** end of session written feedback, pre- and post-course questionnaires, observation in the classroom, interviews, concept maps, teacher portfolios, written assignments, tests/examinations, trainee journal writing.

**The areas the course had most impact on:** attitudes to professional development, beliefs about learning, teaching, classroom management skills, research skills, knowledge of ELT theory, planning skills, sensitive learners.

**MAIN HALL**

**Hong Kong’s Language Proficiency Assessment for Teachers (LPAT): an impact study**

*Neil Drave, Hong Kong Examinations and Assessment Authority*

In the late 1990s, the Hong Kong government made the decision to assess the language proficiency of all serving teachers of English and Putonghua in the public school system, and to make achieving a minimum, ‘benchmark’ level on the assessment mandatory for taking up a permanent teaching position. The test which was developed for this task, the Language Proficiency Assessment for Teachers (LPAT), was therefore designed to be a high-stakes assessment with considerable gate-keeping power. The first LPAT was administered in 2001.

The original decision to ‘benchmark’ serving teachers’ language proficiency garnered considerable criticism from teachers and teacher educators, and there were even street protests against it. Even now, more than ten years and 15 iterations later, and with most candidates being prospective (rather than serving) teachers, the Assessment is not universally liked or understood.

This paper reports on research into the image of the LPAT English Language (LPATE) and its impact on the education sector in Hong Kong, using data from English language media reports. The paper discusses whether the LPATE is considered to have achieved its objectives and how it has been perceived by its various stakeholders. Has the Assessment been a good or a bad thing? What lessons have been learnt about assessment and the formulation of education policy? What has been the nature and quality of media coverage, and has the media served its purpose as a ‘public sphere’ for discussing educational issues? The paper concludes with suggestions for further impact-related research.

**PARALLEL PAPER PRESENTATIONS**

**BARBAROS A**

**A task-based approach to investigating the language proficiency needed to teach EFL**

*Sultan Turkan and Spiros Papageorgiou, Educational Testing Service, USA*

Little is known about the domain of linguistic competence needed to teach English as a foreign language (EFL) at high proficiency levels and how to assess such competence, given the interaction between language ability and specific content knowledge (Elder, 2001). Elder points out that the language domain 1 teachers encompasses general language ability as well as a range of instructional competences required for classroom instruction. This paper aims to disentangle the domain of language proficiency needed to teach EFL effectively at high proficiency levels by examining what is involved in these teachers’ tasks of teaching EFL. More specifically, we present 1) what tasks of teaching identify a group of EFL teachers with high levels of proficiency in English, 2) how these teachers use language to enact these tasks and we infer what linguistic competencies teachers need to have to perform the identified tasks of teaching EFL.

Tasks identified through a review of the relevant literature were included in online logs under four modalities. Five teachers with high proficiency scores on TOEFL were recruited to complete logs. These teachers were chosen from two regions to represent diverse teaching practices and were asked to describe how they enacted language functions and tasks of teaching. In total, 85 log entries were collected during a two-week instructional period. The results of the study offer implications for conceptualizing and assessing the linguistic competence needed to teach EFL at high proficiency levels and using tests that assess such competence in the political context of EFL teaching.

**BARBAROS B**

**Torn in two directions? Functional failure in dual function high stakes exams**

*Tania Horak, University of Central Lancashire, UK*

This research investigated the impact of the Skills for Life Strategy (SfL) (2001) on assessment practices in ESOL teaching in England, and whether these assessments resulted in any washback. The research, undertaken entirely using qualitative methods, found that assessment became considerably more standardised, focussing on a range of external exams, although effects of internal assessment measures were also noted.

While washback was indeed detected, e.g. in staff-student relationships, ‘double accounting’ of students preparing for SfL exams as well as other exams, and to some degree an increased focus on accuracy in classroom work, the washback was not particularly strong. It was also noted to be differential, namely, it was not uniform across the sites studied.
Factors to explain this were investigated, including the variability of stakes for various stakeholders, features of the teachers themselves, and the quality and nature of the communication of the changes in assessment procedures. Findings suggested distortion of the aims of the assessments, caused by perceived pressure to reach targets for securing funding. One major outcome reiterated previous findings by other researchers that when an exam has two functions, e.g., certification and accountability as in the case of SfL exams, if at least one entails high stakes, neither function will adequately be fulfilled.

The results suggested considerable variability, leading to the conclusion that washback studies, which are vital for monitoring exams, need to avoid being simplistic and thus missing key factors which illuminate contextual detail. The nature of washback can easily be masked by superficial investigation.

**15.55-16.30 COFFEE BREAK**

**16.30-16.50 PARALLEL WORK-IN-PROGRESS PRESENTATIONS**

**MAIN HALL**

**The impact of using graphic organizers to test reading comprehension**

*Marci Nelson Özer, Bilkent University School of English Language, Ankara, Turkey*

This presentation aims to reflect on the impact of testing reading comprehension through the use of graphic organizers. At Bilkent University School of English Language (BUSEL), new university entrants work to improve their English before embarking on their academic studies in their departments. They are assessed in part through high stakes, institutional tests which are designed to be highly standardized to increase reliability and efficiency of marking. This has often meant an adherence to straightforward task types such as open-ended short answer questions and 4-option multiple choice questions. However, according to Alderson (2000) “Any single technique for assessment will necessarily be limited in the picture it can provide... and therefore we should always seek to use multiple methods and techniques.” Graphic organizers are an especially effective alternative because they require readers to process texts at a macro level, focusing attention on the knowledge structures that all coherent texts are built on (Grabe, 2009), in addition to assessing understanding of how main ideas are supported and developed. This presentation will first describe how graphic organizers have been used in BUSEL tests, and then examine the positive and negative impacts on exam standardization and marking, the test production process, and teaching. Finally, based on the impacts observed so far, ideas for further development of their use will be touched upon.

**BARBAROS A**

**Monitoring the effects of exam renewal: teachers’ views on preparation material**

*Jayanti Banerjee and Barbara Dame, CaMLA, Michigan, USA*

This study presents the initial findings from impact research on the Examination for the Certificate of Proficiency in English (ECPE) speaking test. A revised ECPE speaking test was launched in 2009. The new speaking test is a paired decision-making task involving two test takers and two examiners. Alongside the launch of the revised test, a validation program was initiated. An initial study provided preliminary evidence of the broader construct of the new speaking test, showing that it elicits a wider variety of communicative functions. A follow-up study indicates that the new ECPE speaking test elicits a greater variety of formulaic expressions, additional evidence of the proficiency level of the test takers. The work in progress to be reported here explores the connection between these findings and the test preparation classroom and textbook materials. With the aim of investigating the impact of the test revision upon the test takers’ preparation and language learning, the current study examines the relationship between the aims of the exam, the language that is elicited, and test preparation (both in classes and at home). The presentation will briefly describe the new speaking test format and the design of the impact study. It will then present findings from an analysis of major test preparation textbooks showing the connections between those materials and the language elicited by the exam. Finally, it will outline the second empirical phase of the project – gathering data on test preparation practices at home and in the classroom.

**BARBAROS B**

**The impact of standardised testing on the teaching of English to teenagers**

*Claudia Mewald, University College of Education Lower Austria, Austria*

This paper presents the first two parts of a longitudinal study on the impact of the implementation of educational standards and standardised testing on the teaching, testing and learning of English as a foreign language in Austrian lower secondary schools. Based on a quantitative on-line survey in 2009, the year of the legal enactment of educational standards and the implementation of a baseline-study focusing on the learners’ output after eight years of foreign language education, the former status quo of the teaching of English is fleshed out and will be used to
describe the educational change until the implementation of the first standardised test in 2013. The data for the description of change derive from non-participant classroom observation in approx. 70 classes at 30 schools, quantitative questionnaires and qualitative interviews with teachers, head teachers as well as change agents from organizational institutions.

Findings from triangulation are presented together with an analysis of the current practice of testing in EFL classes and a comparison with a study on testing in lower and upper secondary schools carried out in the 1990s. The focus will be on the intended effects of educational standards and standardised testing as suggested by the organizational institutions and the actual effects on the concrete classroom practice as suggested by data from observations and interviews. Finally, an outlook will be given on the research design for the next two years which has to take additional impact through the organizational change in general secondary education into account.

16.55-17.15 PARALLEL WORK-IN-PROGRESS PRESENTATIONS

MAIN HALL

Game Over? – Investigating the impact of ECL exams on successful candidates
Róbert Märcz and Gábor Szabó, ECL examinations, Hungary

Effectively any form of language testing is supposed to generate impact that serves the broader goal of furthering candidates’ language learning. Language examinations, however, are often perceived as a goal in themselves, rather than a form of feedback in the course of the language learning process. This is particularly so in the Hungarian educational policy context, where nationally accredited language examinations have great currency. Candidates can earn extra points for university entrance by passing these exams, and nationally accredited language examination certificates are a requirement for university graduation. Accordingly, it is not obvious that candidates, once they have earned their certificates, still continue to develop their language proficiency. In order to gather empirical data on this issue, ECL Examinations have launched a longitudinal project to monitor what the impact of ECL exams is on candidates’ future language learning. The research design is based on periodically sending questionnaires to and conducting on-line interviews with ECL candidates who have successfully passed ECL examinations in different languages at various levels to find out about what candidates do after earning their exam certificates. The presentation intends to demonstrate how this task has been accomplished. Also, the first actual results of the project will be presented, along with an analysis of the responses with special regard to any potential relationship between test impact and test level as well as the language in which the exams have been taken. It will also be investigated whether there might exist a link between test performance and test impact.

BARBAROS

Examining washback in multi-exam preparation classes in Greece
Irini Papakammenou, Agrinio, Greece

Exams have not only become a big business (Spolsky, 2008) but they also play a powerful role in education, politics and society in general (Shohamy, 2008). Thus, more exams are developed and enter the market (Dunkley, 2010) changing the EFL examination market significantly especially in countries with strong examination culture such as Greece (Gass& Reed, 2011; TsagaridPapageorgiou, 2012). Currently, 15 English language examinations are administered at B2 level (Council of Europe, 2001) in Greece so students often participate in more than one exam in the same exam period. The term ‘multi-exam’ class best describes the variety of exams which are taught in the same exam preparation classes in Greece presenting an interesting pedagogical ecology. Such a context presents a challenge to language testing researchers especially with regard to the methodology teachers use to cope with such classes. In addition, there is only related evidence rather than direct research on the influence of tests on both how teachers teach and their attitudes towards the methods they use (Cheng & Curtis, 2012).

This case study, still in progress, aims to investigate the nature and type of the methodology teachers use and the extent to which their approaches are influenced by the nature and requirements of multi-exam classes. The presentation focuses on the theoretical underpinnings by drawing a distinction made between ‘methods’, ‘activities’ and ‘tasks’. As well, the overall research methods, video recorded observations, pre-, post- and follow-up interviews with teachers and the methods of analysis will be presented.
Listening tests - the impact of mode of delivery on performance

Eli Moe, Rønnaug Totland and Ida Arnoldsen Larsen, Vox, Norway

Listening competence has to be captured indirectly, either through speaking or writing. Basic adult L2 language learners (A1 and A2) are a mixed group, including both learners with and without automatic reading skills. Teachers claim that many weak readers are not able to do well on A1 or A2 listening tests because answering questions involves a lot of reading. If this is true, written questions have an unintended impact on some test takers' ability to show what they understand.

The aim of the study has been to check how students and teachers react to different modes of delivering questions in connection with A1 and A2 listening tasks prior to developing a computerized listening test. The study is based on two sets of data: piloting data from two listening tests, both tests in two versions – with and without written questions, teachers’ assessment of their students listening competence (CEFR) as well as data from a questionnaire distributed to teachers of the students involved in the study. Approximately 500 L2 learners and 60 teachers participated in the study. The students sat for two listening tests. They were randomly assigned to one of the two tests. In their first test questions were only presented aurally. In the next round they were sent to the other test, where the questions were presented both aurally and in writing.

Preliminary results show that the students manage both modes of delivering questions, and that none of them react negatively to questions being presented only aurally.

Making the most of the CEFR: developing an assessment programme beyond Europe

Kathryn Brennan, Liverpool International College, UK

The Al Rowdah Academy, UAE, is a partnership between the Institute of Applied Technology (IAT) and Kaplan International Colleges (KIC). Since the inception of this partnership, the UK based assessment team has provided assessments for its language programme. The objective of the programme is to test students within the CEFR level range for language proficiency, A1 – B1+.

This presentation will catalogue the development of the assessment programme from its inception to the present time. It will detail how, by referring to the Common European Framework for Reference, the assessment team were able to utilise the transparency and diversity of the CEFR to arrive at bespoke culturally-appropriate assessments. This approach to assessment writing ensures that the principle and structure of the English language assessment programme and curriculum are informed and guided by the CEFR values and model of language teaching and learning, while at the same time delivering assessments that reflect more closely the socio-cultural needs of the Emirati student group they aim to assess. These assessments provide Emirati students access to tasks designed and set within a culturally-appropriate context which other UK and European tests do not. They also consider the plurilingual aspect of the UAE and its diverse population.

The presentation will highlight that the transparent nature of the Common European Framework lends itself to the development of language programmes beyond a European socio-political context while at the same time providing a ‘globally’ recognisable proficiency scale for language testing and teaching, which is both innovative and student-focused.

Computerized ToEFL exam acceptance

Kubra Karakaya Özyer, The University of North Carolina, USA
Beyza Aksu, The University of Chicago, USA

Standardized testing for college admissions has seen extraordinary growth over the past century and appears to be on the cusp of still more far-reaching changes. For instance, Toefl has been changed dramatically during the past decade and it turned out to a computerized format. Even though there are many validation studies for Toefl as an admission test, there is lack of research that explores examinee's perspective and acceptance toward it. Generally, qualitative research was used to understand the perspective and acceptance of admission tests Walpole, McDonough, Bauer, Gibson, Kanyi, &Toliver, 2005; Mupinga & Mupinga, 2005), but we believe that some type of quantitative research can help researchers to understand how and why students feel about standardized admission tests. To do so, we will explain a computerized Toefl acceptance model and show how understanding of this model will improve our knowledge about Toefl.
The impact of automated measurement of text characteristics

This presentation reports research on automated text analysis systems that analyze edited texts, student writing, speeches, and conversations. The primary focus will be on a system called Coh-Metrix (http://cohmetrix.com and tea.cohmetrix.com), which analyzes texts on multiple measures of language and discourse that are aligned with multilevel theoretical frameworks of discourse comprehension and production. Several dozen measures funnel into five major factors that systematically vary as a function of types of texts, e.g., narrative versus informational) and grade level: narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep (causal) cohesion. We have recently examined a composite measure called formality, which increases with low narrativity, syntactic complexity, word abstractness, and high cohesion. The assignment of texts to students is one of the central concerns of teachers, principals, superintendents, and other experts in educational policy. Students sometimes need to be challenged by assigning texts on difficulty levels that aggressively push the envelope on what they can handle. At other times students need a self-confidence boost by receiving easy texts they can readily comprehend. An alternative text assignment policy would assign texts that are not too difficult or too easy, but at an intermediate zone of difficulty. Yet another policy is to have a balanced diet of texts on the difficulty dimension, with adequate scaffolding for the difficult texts. Coh-Metrix has also been used to analyze student writing and conversation. However, other automated systems have been developed that are more appropriate for these registers of discourse that have minimal or no editing and that tend to have many misspelled words, ungrammatical sentences, and low cohesion. These alternative systems include AutoTutor, Operation ARA, and Writing-Pal in addition to the essay graders that are being used at Pearson Education and Educational Testing Service. The presentation will discuss the grading of writing and conversations in addition to the grading of edited texts.

Art Graesser is a professor in the Department of Psychology and the Institute of Intelligent Systems at the University of Memphis and is a Senior Research Fellow in the Department of Education at the University of Oxford. His primary research interests are in cognitive science, discourse processing, computational linguistics, and the learning sciences. He is current editor of Journal of Educational Psychology, past president of several societies (including Society for Text and Discourse, and Artificial Intelligence in Education), and a member of OECD expert groups on assessment (PISA and PIAAC on problem solving). He received a major award from the American Psychological Association on Distinguished Contributions of Applications of Psychology to Education and Training.

A national exam’s washback on reading assessment in the secondary classroom

Doris Froetscher, Lancaster University, UK; BIFIE, Austria

This paper is relevant to the theme of EALTA 2013 as it reports on a study investigating the effect of a standardized national school-leaving exam introduced in 2008 on how assessment in the classroom has changed as a result. It specifically focuses on the washback of the exam’s reading paper on the assessment of reading in classroom-based tests.

To investigate this, more than 100 class test tasks from the periods pre and post the introduction of the new exam were collected. These tasks, as well as past papers from the new exam, will be analysed regarding their text, task and item characteristics. A qualitative task analysis instrument was specially designed for this study and has yielded high inter- and intra-rater-agreement in pilot studies. First global results indicate that there exist clear differences between pre- and post-2008 class tests: foreseeable ones such as a change towards test methods used in the national exam, but also counter-intuitive changes such as a significant drop in the presence of reading tasks post 2008. It will be interesting to see how the test task characteristics as captured by the instrument are affected by
This paper reports on a pre-test/post-test study in which 30 international students took an eight-week intensive English language course designed to prepare them for a high-stakes EAP test. The students took the test twice: once during the first week of the course and again during the last week. The difference in test scores between these two administrations could therefore serve as a measure of progress.

The focus of the present paper, however, is the findings of a parallel qualitative study which aimed to assess the impact of the test preparation course on test takers’ learning strategies and their own perceptions of their progress. All test takers were interviewed during the penultimate week of the course. They were asked to comment on aspects of the test preparation course and the course materials, and to assess their own progress. They also completed a questionnaire in which they were asked to rate their degree of agreement with a list of statements about their approach preparing for the test. Teachers were also asked to comment via questionnaires on the progress made by individual students. Responses to the interviews and questionnaires were analyzed with a view to ascertaining:
- to what extent test takers consciously distinguish between the goals of language improvement and test preparation;
- which learning strategies test takers perceive to be most effective in relation to these goals;
- how test takers perceive the effectiveness (or otherwise) of the activities and materials which comprised the test preparation course.

The washback of new high-stake examinations in English in Poland
Elzbieta Zawadowska, The Linguistic Academy of Warsaw, Poland

Multifaceted nature of washback triggers critical discussions on its range meaning and character. Undoubtedly it depends not only on the nature of the exam but also on situation in which it is employed and that is why it needs to be investigated for each and every high-stakes examination. This is no less true of the school leaving examination Nowa Matura (NM) introduced in Poland in 2005. This paper will look both at the teacher’s and students’ perceptions of NM and both its positive and negative influence on teaching and learning process. The conclusions are based on quantitative and qualitative research made in 2010 in Poland according to demographic criteria (big cities, towns and villages). The research contained an extensive questionnaire for students and semi-structured interview with their teachers about the same aspects of the school leaving exam. A total of 202 questionnaires and 12 interviews were analysed.

The results of the study should provide a better understanding of the impact of NM on the process of teaching and learning in Poland as well as suggestions for how the exam may be changed to minimize its negative influence on all the participants of the process of education and reinforce its positive impact. Comparisons made across teachers and students representing different teachers areas as well as comparisons made between opinions expressed by teachers and students might help to improve teaching and learning process in all areas of the country.

The impact of classroom assessment on Thai EFL undergraduate students
Jirada Wudthayagorn and Antikar Rongsa-ard, Chulalongkorn University, Thailand

Thai students have regularly undergone various forms of classroom assessment which include multiple-choice, sentence-level and paragraph-level writing which are designed based on teachers’ decision. However, less is known about the physical and psychological impacts of the use of such forms of assessment experienced by the students.

This study aims to report students’ perspectives on these forms of assessment and to present the most preferred methods of classroom assessment as voiced by the students. The participants were 415 first-year Chulalongkorn undergraduate students from 11 faculties who took an English Foundation course. A 5-point Likert scale and open-ended questionnaire was used to collect the data, which was then analyzed using descriptive statistics and content analysis. The findings indicate that summative
formative assessment caused the students’ more anxiety, stress and fear which resulted in such physical symptoms as tiredness, headache, and insomnia. An in-depth analysis also revealed that students’ negative feelings occurred before and while taking the test, and even after completing the test. These students suggested that apart from traditional quizzes and homework assignment, class participation and attendance, and students’ learning progress through time should also be given credits. They realized that they learned something from the course, yet such things or pieces of knowledge might not be covered in the threatening summative evaluation. It seems that the so-called “embedded formative assessment” may play a more vital role in minimizing student negative experiences. Detailed discussion of the findings as well as pedagogical and research implications are also provided.

**Item development, the CEFR, and the perils of Cinderella testing**

*Thom Kiddle, Norwich Institute for Language Education, UK*

This paper looks at three areas of language assessment and draws them together to reflect on their combined impact on test construction. The first area is the creation of a new item type, which aims to measure candidates’ understanding of textual cohesion in reading processes, through what is referred to as intra-text-banked gap-filling. The aspects of the process of reading which underlie the construct of this item type are not in harmony with the comprehension-outcome dimensions of reading as reflected in the Common European Framework of Reference descriptors for Reading Comprehension. This has major implications for the inclusion or omission of such item types in construction of tests which need (for political or practical purposes) to be related to the CEFR. The dangers inherent in this are underlined with reference to ‘Cinderella testing’ – the idea that in attempting to relate to external frameworks, the content and construct of tests (in their ability to cover process and products aspects of language ability) is threatened – c.f. Prince Charming’s approach to finding his future princess merely by measuring shoe size! The paper concludes with a call for research into how stakeholders in proficiency test construction and use can develop approaches to adapting such frameworks as the CEFR to take into account test instruments which assess the aspects of language processing as well as the products, and henceforth instruments which promote classroom attention to processes of language comprehension.

**Developing and validating band levels for reporting overall examinee performance**

*Xiaoming Xi, Spiros Papageorgiou and Youngsoon So, Educational Testing Service, USA*

For a test to be useful in decision-making and to exert a positive impact, the meaning of scores should be transparent to its users. One way to achieve transparency is through the use of band levels and performance descriptors (Alderson, 1991). Despite the growing practice of linking test scores to external language proficiency frameworks (e.g., CEFR) when interpreting scores, little is known about the development and validation of the reporting levels and the descriptors that are closely aligned with the content and intended uses of a test (Garcia et al., 2007). This study aimed to develop band levels and accompanying descriptors for reporting overall examinee performance on a test of English as a second/foreign language. Three questions were addressed:

- How many levels can be meaningfully distinguished?
- How reliable is the classification of students into these levels?
- How should performance descriptors be developed to provide meaningful information to score users?

The performance data from 2,931 students who took the test were used. The reliability of classification decisions was estimated using RELCLASS (Livingston & Lewis, 1995). In order to construct meaningful descriptors for the band levels, multiple sources of information were examined, including scoring rubrics, characteristics of items, and typical student performance profiles. Feedback was incorporated from teachers, school administrators and other users from five countries in finalizing the score reporting scheme. We stress the importance of establishing the psychometric quality of band levels and the empirical basis for performance descriptors, and discuss the implications for similar efforts.

**Examining the consequential validity of a standardised test in a classroom setting**

*Christine Doe, Mount Saint Vincent University, Canada*

There has been a long-standing debate about how consequences stemming from test use should be included in the validation process. This paper draws on consequential validity data to provide evidence for a validity argument making a case for the use of a standardised test for formative purposes in a classroom setting. The Assessment Use Argument (AUA) (Bachman & Palmer, 2010) was employed to
The Impact of Language Testing and Assessment

12.30-12.55 PARALLEL PRESENTATIONS

MAIN HALL

Exploring the criterion-related validity of the new HSK and its impact
Yang Lu, University of Nottingham, UK

It was reported in 2011 that Mandarin Chinese has 100 million learners. Many students in U.K. universities have studied Chinese as part of their degrees or optional modules. Apart from the institutional assessment, there has been a great demand for standardised tests for proficiency levels in Chinese and the New HSK has become the most widely acknowledged test with an annual candidature of more than 100,000. The test also claims a “test-teaching correlation” and to promote learning through testing.

For U.K. HE, therefore, it has been necessary to know how useful, meaningful and appropriate the New HSK test is to their students and Chinese programmes by relating their benchmarks and assessment to those by the standardised test. This paper attempts to explore the criterion-validity of the New HSK at the threshold level, Level 3 and investigates the agreement between learners’ performance on the institutional assessment and that on the New HSK exam.

The criteria and papers of the New HSK Level 3 and the U.K. university are introduced and compared. Scores awarded by the New HSK paper to the subjects and by the achievement assessment are examined to see if they correlate and the concurrent and predicative validity of the New HSK Level 3. Findings suggest a great discrepancy between the two measures, especially in written language. Discussion on the impact, possible political involvement and dilemmas in testing written language in Chinese will be given.

BARBAROS B

Fairness in the secondary schools: A Singapore case
Chih-Min Shih and Li-Yi Wang, Nanyang Technological University, Singapore

Fairness has drawn attention from researchers especially in the past two decades. Both Language Testing and Language Assessment Quarterly have a special issue on fairness. However, most relevant literature on this subject is theoretical-oriented, and very few studies have focused on how fairness is implemented in the classroom. The present study intends to fill this research gap, examining how secondary school teachers of English in Singapore conceptualize and practice fairness when they conduct classroom assessment. 40 secondary school teachers with diverse backgrounds were interviewed. 5 out of the 40 interviewees were chosen for classroom observations. The results show that teachers applied many strategies to make their tests fair to students, including consulting senior colleagues, relating tests to their teaching, using culturally-relevant materials, standardizing marking criteria across department, specifying bands descriptions to students, and controlling variables in testing environments. In addition, most teachers claimed they would consider students’ cultural, linguistic, and ethnic backgrounds when they designed or wrote a test paper for the purposes of improving fairness and engaging students better. Only a few argued that considering student characteristics might jeopardize the objectivity of tests and the preparation for high-stakes exams. Also, only a few would give special considerations to students who have just arrived in the country or who are from socio-economically disadvantaged families when they assign a score to these students. On the other hand, most teachers claimed they would not give special treatment to these students in their grading practice because of school policy or/and their professional decisions.

12.30-12.55 PARALLEL PRESENTATIONS

MAIN HALL

Exploring the criterion-related validity of the new HSK and its impact
Yang Lu, University of Nottingham, UK

It was reported in 2011 that Mandarin Chinese has 100 million learners. Many students in U.K. universities have studied Chinese as part of their degrees or optional modules. Apart from the institutional assessment, there has been a great demand for standardised tests for proficiency levels in Chinese and the New HSK has become the most widely acknowledged test with an annual candidature of more than 100,000. The test also claims a “test-teaching correlation” and to promote learning through testing.

For U.K. HE, therefore, it has been necessary to know how useful, meaningful and appropriate the New HSK test is to their students and Chinese programmes by relating their benchmarks and assessment to those by the standardised test. This paper attempts to explore the criterion-validity of the New HSK at the threshold level, Level 3 and investigates the agreement between learners’ performance on the institutional assessment and that on the New HSK exam.

The criteria and papers of the New HSK Level 3 and the U.K. university are introduced and compared. Scores awarded by the New HSK paper to the subjects and by the achievement assessment are examined to see if they correlate and the concurrent and predicative validity of the New HSK Level 3. Findings suggest a great discrepancy between the two measures, especially in written language. Discussion on the impact, possible political involvement and dilemmas in testing written language in Chinese will be given.
BARBAROS A

Using and interpreting English proficiency test scores in the graduate admissions process

April Ginther, Purdue University, USA
Catherine Elder, University of Melbourne, Australia

Current theoretical discussions of test impact and the consequential aspects of test validity have led to increased interest in how test scores are used and understood by stakeholders within particular domains. Adopting an instrumental case study approach, this study examines levels of knowledge about the English language tests (TOEFL, IELTS, and PTE) used for selection in two academic contexts and the uses of test scores in local decision-making by graduate faculty. Data for the study were gathered via an online survey and follow-up interviews probing the basis for participants’ beliefs, understandings, and practices.

The presentation focuses on the results of the 50-item survey completed by 232 respondents at a large Research 1 U.S. university, 246 respondents at a large Research 1 Australian university, and 30 follow-up interviews at both institutions. Responses reveal that English test scores, once entry-level requirements are met, tend to have very limited impact on admissions decisions as compared to evaluations based on other kinds of available evidence. In both contexts, respondents emphasized (1) the importance of English for academic success; (2) dissatisfaction with current levels of English among graduate students; (3) limited knowledge about or understanding of the major English tests used for selection; and (4) a concomitant lack of preference for any one of the 3 tests accepted for admissions purposes. Conclusions discuss the validity implications of findings and the challenges of ensuring that critical information about language tests is appropriately tailored for different audiences given the complex mechanisms and distributed responsibilities for student selection.

BARBAROS B

Peer evaluation impact on testing oral presentation skills

Okan Önal, Kuleli Military High School, Marmara University ELT Department, Istanbul, Turkey

This study investigates the reliability and advantages of peer evaluation in testing oral presentation skills. The study was conducted on 32 first-year university students at Marmara University ELT department in 2012 Fall Semester. The participants’ oral presentation performances were evaluated by the researcher using a 12-item scoring rubric. In the meantime, each participant was evaluated by their classmates using the same criteria. The evaluation scores for 32 students were later compared and correlated in terms of both individual items and final scores. The results showed considerable consistency between peer evaluation and researcher evaluation particularly in the performances of high-achievers. What is more, peers demonstrated significant reliability in particular criteria items. Finally, participants stated that evaluating the performances of their classmates not only increased their awareness towards better oral presentation skills, but also helped them improve their own performances.

12.55-14.30 LUNCH AND POSTER SESSIONS

14.30-15.30 PAPER PRESENTATIONS | MAIN HALL

14.30-15.00 Demonstrating validity for high-impact language tests

Nivja de Jong, Utrecht University, The Netherlands
Jelle Goeman, Leiden University Medical Center, The Netherlands

Language tests have high impact because they discriminate between people and this discrimination is subsequently used to decide on people’s future. Therefore, language tests should measure what they claim to measure. Borsboom et al. (2004) argue that a test is a valid measurement of an attribute if variations in the attribute causally produce variation in measured outcomes. This means that, (1) test developers should use item characteristics to construct items that differ in difficulty in a theoretically predictable way, considering the predicted impact of the test; (2) items that are assumed to differ in difficulty with respect to the attribute should measure a unidimensional construct; and (3) found item difficulty should be related to the item characteristics as defined in (1) only.

This straightforward conception of validity has rarely been used for high-impact language tests. We will show that it can be surprisingly simple to demonstrate validity using Borsboom’s criteria. As an example, we take a productive vocabulary test which was performed by 257 participants. The test
consisted of 90 items in which target words that were omitted from meaningful sentences had to be filled in. Target word frequency was beforehand identified as the item characteristic that should be related to item difficulty. To confirm unidimensionality, we used a new measure, related to Cronbach’s alpha, but without its known problems (Sijtsma, 2009). Finally, we show that, as predicted, actual item difficulty is strongly related to target word frequency, and not to other item-characteristics such as carrier sentence length and complexity.

**An interpretive argument for automated writing evaluation as a formative assessment tool**

*Stephanie Link and Ahmet Dursun, Iowa State University, USA*

In second language classrooms, the use of Automated Writing Evaluation (AWE) has become a promising area of investigation in the United States as a tool for providing immediate feedback on students’ writing (Warschauer and Ware, 2010). The revision process of L2 writing with AWE is grounded in a formative assessment approach, focusing on continued use of feedback (Rea-Dickins & Gardner, 2000). Although the role of AWE has shown potential for the language classroom (Chen & Cheng, 2008), user-centric versus system-centric evaluation of such tools has yet to be discussed in literature.

In this presentation, we provide a framework for evaluating the appropriateness and adequacy of one AWE tool, *Criterion®* by ETS, offering an interpretive argument, which “specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions” underlying student performance and decisions-making (Kane, 2006, p. 23). Our framework implements Kane’s (2006) argument-based approach to validation, concepts from Chapelle, Enright and Jamieson’s (2008) interpretive argument, and Bachman and Palmer (2010)’s claim of consequences to outline the ways AWE feedback should be meaningfully connected to students’ revision processes and teachers’ actions. Our presentation will also highlight data from our multi-semester longitudinal mix-methods research conducted at a mid-west university in the U.S. The data provides evidence in support of several of the inferences and allows stakeholders in the field of language testing to make informed decisions about the use of AWE in situation-specific contexts. We conclude with designs for future research and ongoing AWE validation studies.

15.00-15.30  **COFFEE BREAK**

16.00-17.30  **ANNUAL GENERAL MEETING | MAIN HALL**

19.00-23.30  **SOCIAL EVENT: Bosphorus Cruise and Conference Dinner**
Sunday, May 26th

09.30-11.00 PAPER PRESENTATIONS | MAIN HALL

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>09.30-10.00</td>
<td>What is the impact of diagnostic language tests?</td>
<td>J Charles Alderson, Lancaster University, UK, Ari Huhta, Lea Nieminen, Riikka Ullakonoja, Eeva-Leena Haapakangas, University of Jyväskylä, Finland</td>
</tr>
<tr>
<td></td>
<td></td>
<td>This paper brings together three strands of research in language testing and applied linguistics: washback, diagnostic testing, and feedback. We first briefly discuss the concept of washback and research into its nature and the role of teachers and students in bringing it about, and show that the belief that there is a direct relationship between the characteristics of a test and its impact on classroom practices is too simplistic. We then present an overview of recent developments in diagnostic testing and argue that such tests have the potential to increase and systematize the provision of feedback to language learners and, consequently, to increase the positive washback from tests. Indeed, one of the defining characteristics of diagnostic tests is that they are designed to have a positive effect on learning. We report on research related to two projects on diagnosing second or foreign language (SFL) proficiency. The first project was the development of the on-line diagnostic assessment system DIALANG. The second is an ongoing interdisciplinary study into the prediction and diagnosis of reading and writing in a SFL. We synthesize findings on the use of DIALANG feedback and discoveries from the more recent research on predicting difficulties in SFL learning that shed light on what feedback from diagnostic tests could be like and how such feedback could impact learning and teaching. Finally, we discuss some potential issues with the impact of diagnostic tests, such as their focus on specific details of language and proficiency.</td>
</tr>
<tr>
<td>10.00-10.30</td>
<td>Task design for automatic scoring of interactive spoken pragmatic skill</td>
<td>Jared Bernstein, Mallory Klungtvedt, Kayla Neumeyer, Elizabeth Rosenfeld and Alistair Van Moere, Pearson</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A principled objection to computer scored tests is that they may lead to unduly narrowed instruction and learning. Machines can elicit spoken responses to be scored by computer with high accuracy if the expected responses are sufficiently constrained. Efficient and accurate speaking tests have implemented tasks like short questions, elicited imitation, and spoken summaries, from which syntax, lexis, and phonology can be scored with precision. However, aspects of proficiency like sociopragmatic competence may operate over multiple turns and require instruction to shift toward demonstrations of pragmatic knowledge in use (Kasper and Roever, 2005). We present the development and evaluation of three sets of tasks that elicit spoken turns that expressing pragmatic intent, but which should be machine scorable. The first task type, which has been piloted on school children (n=710, aged 5 to 12; approximately half English language learners), elicits (by speech and figure) a single spoken turn to accomplish a situated goal. These items are being trialed on adults (n=210), along with a second type that elicits two turns within a single situation, each turn probing different aspects of pragmatic competence. Performances on both tasks are scored on judged social effectiveness, with secondary credit for pragmatic tone and surface features. Results will explain 1) how pragmatic ratings differ from general proficiency, and 2) how the single- and dual-turn elicitions compare. Finally, the impacts of such pragmatic performance tasks are discussed, with particular emphasis on how such an automated assessment could be used in classroom and computer-mediated language learning.</td>
</tr>
<tr>
<td>10.30-11.00</td>
<td>Factors affecting long-term examination impact and the fate of the examinations themselves</td>
<td>Dianne Wall, Trinity College, London, UK</td>
</tr>
</tbody>
</table>
|       |                                              | How much impact can new examinations be expected to have, both in terms of classroom washback and effects on the educational context? Even if the impact is judged to be extensive and positive, can we
reasonably expect examinations to endure intact, or must they change, or even disappear, in response not necessarily to developments in the way we view language proficiency but to social and political pressures in the environment? This paper presents a long-term view of examination impact, taking as starting point the aspirations of teams designing new school-leaving examinations in countries that gained independence in the early 1990s. It describes the challenges the teams faced in introducing new forms of assessment and the skills and mechanisms they had to develop in order to convince stakeholders that their examinations represented an opportunity rather than a threat to language teaching. It discusses the challenges that appeared in the years that the examinations became the ‘status quo’, and the question of how to balance academic considerations such as curriculum intentions and assessment validity and reliability with real-world issues such as limited resources, competing priorities and politics at various levels. The paper represents an analysis and interpretation of documents written in the early days of examination reform in four countries of the former Soviet Union, and interviews with examination designers, teachers, and policy-makers 10 to 15 years after the introduction of examinations that were meant to both reinforce and influence innovations in language teaching.

11.00-11.30 COFFEE BREAK

11.30-12.30 PLENARY 3 | Dan Douglas, Professor Emeritus
Iowa State University, USA (dandoug@iastate.edu)

MAIN HALL | Nobody Speaks English Here Today: The Notion of Impact in the Context of Aviation English

After discussing the concept of impact in second language assessment and noting the general absence of the notion in many professional codes of ethics and practice, I will elaborate on impact in the context of aviation English. I will focus particularly on current International Civil Aviation Organization language proficiency requirements which state clearly that they refer to native and non-native speakers of English alike; yet ICAO tests only the latter group of pilots and air traffic control officers, and in doing so, ignores a number of factors deemed necessary for international air safety. These factors include issues of aviation experience and expertise, lack of professional knowledge by either pilots or air traffic controllers, lack of adherence to standard conventions of phraseology, and a shared responsibility for misunderstandings and for communicative repairs. I will argue that the impact of such policies means not only that non-native English speaking pilots or air traffic controllers who do not show the minimum required level of English proficiency are barred from practicing their profession but also that native English speaking professionals remain largely unaccountable for communication problems that bear heavily on air travel safety. As a way forward in achieving accountability and fairness in aviation English assessment, I will suggest that a test might include a focus on intercultural communicative abilities for both native and non-native English speakers and I will speculate on what such a test might look like.

Dan Douglas, Professor Emeritus at Iowa State University, holds a Ph.D. in Linguistics from Edinburgh University, and has held teaching and research appointments at the University of Khartoum, Hiroshima University, the University of Michigan, Lancaster University, and the University of London. He is currently President of the International Language Testing Association. His books include Understanding Language Testing ( Hodder Education, 2010); Assessing Language through Computer Technology, with Carol Chapelle (Cambridge, 2006); and Assessing Languages for Specific Purposes (Cambridge, 2000). He was co-editor of the journal Language Testing from 2002 to 2007, and currently serves on the editorial boards of several journals.
Flipped washback? Improvisational theatre as a means of preparing dialogical oral assessment
Karin Vogt and Raphaelle Beecroft University of Education Heidelberg, Germany

Oral exams have become part and parcel of classroom-based language assessment, final and high-stakes exams included. At the same time, foreign language teachers face a challenge in preparing learners, particularly weak ones, for these exams that often include a dialogical component. In the German context, these exam formats have only relatively recently been introduced.

The research project “Initiating, eliciting and assessing speaking competence in the foreign language classroom” tries to turn the notion of washback upside down in line with principles of backward design (Wiggins & McTighe 2005) to eventually create positive washback. It aims at preparing learners for a final exam dialogical task or for formative assessment during the school year, using improvisational theatre as a method to elicit extended and spontaneous discourse in dialogues. The research interest in LTA centres around the question of whether implementing structured improvisational theatre sessions contribute to a higher success rate in final exam dialogical tasks. Three classrooms (n=72) in lower secondary are part of the mixed-methods study that heavily draws on action research methods and consistently involves teachers in in-service teacher training. Fully spelled-out research questions, the methodology of the study as well as first findings of the pilot study will be outlined in this progress report poster presentation.

Tests for Turkish as a heritage language
Sibylle Plassmann et al. – language tests, Germany

In 2008, telc – language tests were commissioned to develop tests for Turkish by the Ministry for Education in the German state of Hesse. The target group were school pupils studying Turkish as a heritage language. The intended impact was defined as follows:
- the introduction of modern methods of language teaching by using a task-oriented test,
- better motivation of learners by setting a clear aim,
- providing school-leavers (especially low achievers) with a valuable certificate in order to facilitate the transition into vocational training,
- and on a macro-level: defining competence in Turkish as a valuable asset, thus giving a political signal to the whole Turkish community and to German society in general.

The telc Turkish test suite for young learners (A2–B2) has been in use for some time now, and first research could be carried out in order to monitor aspects of impact. The poster will show the different stakeholders’ perspectives: learners, their parents and teachers, school authorities and ministries as well as media uptake. Even at this early stage, some aspects of impact have become quite visible: from the test takers’ raised level of motivation (as documented in a first survey) to increased public visibility of Turkish (cf. high profile ceremonies for awarding certificates).

For the language testers themselves the translation of the CEFR into Turkish has been the greatest challenge. The new “AOÖC” will, however, provide an even more profound impact on the teaching and learning of Turkish than the tests themselves.

Righting writing practices: The impact of exam reform
Kremmel Benjamin, Eberharter Kathrin and Konrad Eva, University of Innsbruck, Austria

In 2004, Austria introduced a new curriculum linking foreign language teaching to the CEFR, and by consequence, initiated a change in assessment practices in its school-leaving exams from a factual knowledge based focus to a competence based approach. The curriculum stipulates equal weighting of all language skills, replacing writing as the predominant skill in classroom teaching and assessment. This study aims at investigating changes in teaching and classroom assessment of writing after the introduction
The Impact of Language Testing and Assessment

'SEnglish' is one of them. This study aimed to find out how comparable the two modes are before the pencil and paper version is abandoned completely. The main purpose of this study is to compare and contrast the pencil and paper and the computer-based versions of the Language Examinations, administered by the Department of English and American Studies at the Faculty of Arts, Masaryk University, Brno. These compulsory examinations are taken by students at the end of their first, second and third years and serve as prerequisites to their further studies. The examinations measure proficiency of the students at advanced and proficient levels of English. Recently, computer-based versions of the Language Examinations have been implemented. Two main areas will be presented, namely Validity Issues and Computer Assisted Language Testing, in order to explain the context of the project. Previous research in the field will be summarised, focusing predominantly on notion of validity and its various aspects. The increasing role of computer-based tests is associated with rapid advances in technology. Technology-related positive and negative aspects, such as consistency, or nervousness respectively, will be considered. The main purpose of this study is to compare and contrast the pencil and paper and the computer-based versions of certain Language Examinations. It attempts to identify possible differences in the two modes of administration, taking into consideration the latest research in the field, test results as well as the individual students’ opinions and feelings. With the implementation of the computer-based tests into the testing procedure, there is an urgent need to find out how comparable the two modes are before the pencil and paper version is abandoned completely.

The impact of gender on interviewer behaviour during oral proficiency interviews

Ene Alas and Suliko Liiv, Tallinn University, Estonia

The poster will present the results of a study investigating the behaviour of interviewers during oral proficiency interview within the framework of national examination in the English language in Estonia and does so from a gender perspective. The study was interested if noticeable differences can be traced in the behavioural patterns of male and female interviewers that could lead to either favouring or placing the candidates in a disadvantaged position during an oral interview, not allowing them to display their language proficiency to the full. The study involved an equal number of male and female interviewers whose interview language was subjected to content analysis in all the stages of the interview. The results showed an overall attempt to follow the scripted interview format envisaged for this examination, but none of the interviewers completely followed the script. Deviations ranged from significant changes like leaving out entire stages of the interview to fairly minute alterations like introducing additional discourse markers. Both men and women used accommodation strategies to help the candidate, yet the nature of accommodation seemed to be gender-specific. Time-management was another area of contention with the interviewers and there, too, male and female interviewers displayed quite distinct strategies. Overall, a set of features could be identified with both gender groups that, if not better managed by further interviewer training, may potentially affect the validity of test results.

Keywords: interviewer, oral proficiency interview, validity, reliability, scripted interview, gender variation, standardisation.

Comparing two modes of test administration

Linda Nepivodová, Masaryk University, Brno, The Czech Republic

The project explores certain Language Examinations, administered by the Department of English and American Studies at the Faculty of Arts, Masaryk University, Brno. These compulsory examinations are taken by students at the end of their first, second and third years and serve as prerequisites to their further studies. The examinations measure proficiency of the students at advanced and proficient levels of English. Recently, computer-based versions of the Language Examinations have been implemented. Two main areas will be presented, namely Validity Issues and Computer Assisted Language Testing, in order to explain the context of the project. Previous research in the field will be summarised, focusing predominantly on notion of validity and its various aspects. The increasing role of computer-based tests is associated with rapid advances in technology. Technology-related positive and negative aspects, such as consistency, or nervousness respectively, will be considered. The main purpose of this study is to compare and contrast the pencil and paper and the computer-based versions of certain Language Examinations. It attempts to identify possible differences in the two modes of administration, taking into consideration the latest research in the field, test results as well as the individual students’ opinions and feelings. With the implementation of the computer-based tests into the testing procedure, there is an urgent need to find out how comparable the two modes are before the pencil and paper version is abandoned completely.

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

Chamnan Para, University of Bristol, Uk

In Thailand, there is the national university examination, O-NET (the Ordinary National Educational Test). Twelve grade students sit this examination to apply for universities. O-NET has total eight subtests and ‘English’ is one of them. This study aimed to investigate the washback of the English test.
The results indicated that the majority of participants were negatively affected. For English teachers, their English materials and lessons were designed to gear toward the test items in the English test of the O-NET. They had great difficulties in maintaining the balance between the compulsory school curriculum and the expected success rate of students in passing O-NET. However, some teachers still strictly taught English to students in line with the assigned curriculum. All four 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 O-NET. 

Regarding students, most students tended to be more selective in the English content they study. Students intentionally drilled the content that was likely to be appeared in the test and disregarded the unlikely ones. In some cases, students even skipped normal classes to attend private tutorial schools in order to have a full test preparation.

It is obviously seen that the English test generates more negative impacts than the positive ones. Moreover, the differences between the school curriculum and those real test items in the O-NET’s English test result in generating the unwanted impacts to English teachers and students.

The impact of the speaking component of an institutional test on students’ anxiety level

Selen Sökmen, Çağ University, Turkey

There is a considerable amount of research on the impact of tests on students’ anxiety levels in the field of language testing and assessment. Test anxiety might cause poor performance because of the stress students feel before, during, and/or after a test. No matter what types of questions the test covers, there is always a high possibility that some of the test takers suffer from test anxiety. This problem might even increase when the test includes a spoken component as it is well known that language learners feel anxious, apprehensive, and nervous while speaking in the target language. The present study has examined the impact of newly-introduced spoken component of an institutional proficiency test on students’ anxiety level. In order to reach this aim, 140 students at a preparatory school of a university were asked to fill in a questionnaire before they took the institutional proficiency exam, which included the spoken component for the first time. The questionnaire revealed whether and to what extent students’ believed they would experience anxiety in the written and spoken part of the exam. In order to investigate how they felt during the spoken test, 40 of the students were interviewed soon after they completed the spoken part of the test. This study is expected to reveal the importance of considering all possible impacts of a test before making any changes.

Exploring Washback Effects of High-Stakes Tests on ELT in Russia

Ludmila Kozhevnikova Viktoria Levchenko

Since the late 1980s the Russian educational system has been going a radical transformation. One of the major innovations of the education reform was the introduction of a new type of the school-leaving exam – the Russian State Examination (RSE) which brought about dramatic changes into the system of educational assessment, both formative and summative.

Another high-stakes exam – FEPO – was introduced at tertiary level. It is the end-of-the course multiple-choice test in each subject. It was introduced with the purpose of improving comparability between students’ academic achievements in higher education area.

Although the RSE and FEPO are high-stakes state - administered examinations, they are not without problems many of which stem from the fact that in Russia, there is no system of formal training in testing. In other words, many people who are involved in testing and assessment either lack assessment literacy or have a very simplistic view of a test per se. To improve the current situation, a TEMPUS project was started in Russia in 2011. The Promoting Excellence in Testing and Assessment of English (PROSET) project aims to provide high quality training in language assessment. A key project goal is ‘to create university modules in language assessment that reflect European quality standards, but are well adapted to the local needs of the Russian

The aim of the postal presentation is to present the case study on washback effect of the RSE and FEPO report on some results of the implementation of the PROSET project in Russia.