13th EALTA ANNUAL CONFERENCE

3 – 8 May 2016 - VALENCIA -Spain

PRE-CONFERENCE PARALLEL WORKSHOPS 3-5 May 2016

CDL- Building 4P

TUESDAY 3 12:30-13:30	**Registration-** (Ground Floor- CDL))			
WORKSHOPS 3-5 MAY	**WORKSHOP 1** Multimedia (1st floor)	**WORKSHOP 2** Room 9 (Ground floor)	**WORKSHOP 3** Seminar(1st floor)	**WORKSHOP 4** American Space(1 floor)
TUESDAY 3 13:30-17:30	**Online Testing Resources & Computer-based Assessment** ***Jenny Liontou***: Greek Ministry of Education ***Dina Tsagari*** : University of Cyprus, Cyprus	**ASSESSING WRITING: Designing better rubrics, building better rating communities** ***Emma Bruce*** : City University of Hong Kong ***Liz Hamp-Lyons*** : Editor of the journal Assessing Writing	**Quality assurance in test development: a hands-on introduction to task specifications, item writing, trialling and statistical analysis** ***Gwen Caudwell***: Aptis Product Development Manager, British Council ***Kevin Rutherford***: Aptis Test Production Manager, British Council ***Judith Fairbairn:*** Test Development Consultant, British Council ***John Tucker:*** Aptis Test Production Co-ordinator	**A practical approach to tackling challenges in Quality Assurance (QA) in test development projects** ***Neus Figueras***: Generalitat de Cataluña ***Elaine Boyd:*** Institute of Education, University College, London
15:00 COFFEE BREAK				
WEDNESDAY 4 10:30-13:30				
15:00 COFFEE BREAK				
15:00-17:30				
THURSDAY 5 9:30-12:30				
11:00 COFFEE BREAK				

SIG MEETINGS

THURSDAY 5 May: 14:00- 17:30

CDL- Building 4P

SIG 1 Room 3 (Ground floor)	Assessing Speaking
SIG 2 Room 9 (Ground floor)	CEFR
SIG 3 Seminar(1st floor)	Special Interest Group 'Classroom-based Language Assessment'
SIG 4 American Space(1 floor)	Assessment of Writing and Assessment for Academic Purposes

EALTA's 13th CONFERENCE- 5-8 May 2016

Assessment of what...? Revisiting the issue of construct(s)

CONFERENCE PROGRAMME

THURSDAY 5th MAY

12:00-18:00	**REGISTRATION** (Ground floor - Language Centre CDL- Building 4P)
18:00-20:00	**WELCOME RECEPTION** (Ground floor - Language Centre CDL- Building 4P)

FRIDAY 6th MAY

8:30-9:45	**REGISTRATION** (Ground floor - NEXUS- Building 6G)		
9:45- 10:00	**OPENING CEREMONY** (Conference room : Ground floor - NEXUS- Building 6G)		
10:00-11:00	**KEY NOTE PRESENTATION** (Conference room : Ground floor - NEXUS- Building 6G) **Constructing Language and Literacy Practices– Implications for Assessment** ***Constant Leung:*** *Professor of Educational Linguistics* Chair*: Cristina Perez-Guillot*		
11:00-11:30	**COFFEE BREAK**		
11:30- 13:30	**PAPER PRESENTATIONS** (Conference room : Ground floor - NEXUS- Building 6G) Chair*: Neus Figueras*		
11:30- 12:00	Language assessment literacy for test developers: towards an understanding of test constructs *Cristina Rodriguez, Julia Zabala*		
12:00- 12:30	"Study on comparability of language testing in Europe": focus on the constructs *Esther G. Eugenio, Michael Corrigan*		
12:30- 13:00	An empirical investigation of the relationship between L2 vocabulary size and L2 listening comprehension at the B1 CEFR-level in English and French *Britta Kestemont, Ann-Sophie Noreillie*, Kris Heylen, Piet Desmet, Elke Peters (*Joint First author)*		
13:00- 13:30	French speakers reading German and German speakers reading French – do they take different tests? *Monique Reichert, Charlotte Krämer*		
13:30– 15:00	**LUNCH BREAK**		
15:00- 16:15	**WORK-IN-PROGRESS SESSIONS** (2nd floor NEXUS- Building 6G)		
	Room 2.7 (2nd floor) Chair*: Jamie Dunlea*	**Room 2.8 (2nd floor)** Chair*: Dina Tsagari*	**Room 2.9 (2nd floor)** Chair*: Marta Genísi*
15:00- 15:25	Editing tasks – (what) do they add to the writing construct? *Sonja Zimmerman, Anika Müller-Karabil*	ECD for MSA: Developing comprehensive construct definition *Bjorn Lennart Norrbom, Abdulrahman al-Shamrani, Yong Lou*	Developing an adaptive diagnostic test using evidence-centered design *Evelyn Reichard, Ingrid Bresser, Wilma Vrijs*

15:25- 15:50	The construct of writing across contexts. *Carol Spöttl, Sonja Zimmerman, Jayanti Banerjee*	25 years of assessment at the EOI (Government-owned Language School) in Malaga. *Carmen Medina.*	Assessment of what ...? Validating a large-scale, high-stakes speaking test *Thomas Koidl*
15:50– 16:15		Towards Localisation of a Test: What are the Essential Considerations? *Ying Zheng, Yanyan Zhang*	Social science reading constructs and the CEFR reading level of textbooks *Barbara Blair, Kari Telstad Sundet*
16:15-16:45	**COFFEE BREAK**		
16:45-17:45	**PARALLEL PAPER PRESENTATIONS** (2nd floor NEXUS- Building 6G)		
	Room 2.7 (2nd floor) Chair*: Carol Spöttl*	**Room 2.8 (2nd floor)** Chair*: Asunción Jaime*	**Room 2.9 (2nd floor)** Chair*: Carolyn Westbrook*
16:45- 17:15	Factors affecting listening: operationalising a listening construct *Julia Zabala, Cristina Perez-Guillot*	Flip teaching and construct irrelevant variance in classroom-based assessment *María Luisa Carrió Pastor*	Defining vocabulary as a unitary construct: a CEFR-based framework *Veronica Benigno, John De Jong*
17:15– 17:45	Defining the listening construct in multimodal environments *MariCarmen Campoy-Cubillo, Mercedes Querol-Julián*	Using a corpus to validate test construct: The case of epistemic markers in the Trinity Lancaster Corpus *Elaine Boyd, Vaclav Brezina*	The construct of the collocation knowledge in language testing *Margarita Alonso-Ramos*
19:00	**VISIT TO THE CITY**		

SATURDAY 7th MAY

9:00-10:00	**KEY NOTE PRESENTATION** (Conference room : Ground floor - NEXUS- Building 6G) **Fluency and Interactivity** ***April Ginther***: Associate Professor; Director of OEPP Chair*: Claudia Harsch*		
10:00- 11:00	**PAPER PRESENTATIONS** (Conference room : Ground floor - NEXUS- Building 6G) Chair*: Claudia Harsch*		
10:00- 10:30	Construct(s) measured in face-to-face and video-conferencing delivered speaking tests *Fumiyo Nakatsuhara, Chihiro Inoue, Vivien Berry, Evelina Galacz*		
10:30- 11:00	Test-taker interaction: How do raters perceive the construct? *Linda Borger*		
11:00- 11:30	**COFFEE BREAK**		
11:30- 13:30	**PARALLEL PAPER PRESENTATIONS** (2nd floor NEXUS- Building 6G)		
	Room 2.7 (2nd floor) Chair*: Cristina rodriguez*	**Room 2.8 (2nd floor)** Chair*: Elaine Boyd*	**Conference room (Ground floor)**
11:30- 12:00	Revisiting an EAP Speaking test: what is 'EAP Speaking', exactly? *Bruce Howell*	The CEFR as construct across national contexts – Is your B2 my B2? *Franz Holzknecht, Ari Huhta*	
12:00- 12:30	The construct validity of a C2 exam for university students	Exploring interactions between learner and task characteristics in a reading	**SYMPOSIUM (90'')** (Conference room : Ground fl)

	Laura Riera Grau	test of French for young learners – the example of the language of rubrics and response *Katharina Karges, Malgorzata Barras, Peter Lenz*	Revisiting the speaking construct: Multiple perspectives **Organizers:** *India C. Plough* (Michigan State University) *Jayanti Banerjee* (Worden Consulting LLC) **Presenters:** *Spiros Papageorgiou* (Educational Testing Service) *Cathy Taylor* (Trinity College London) *Alistair Van Moere* (Pearson) **Discussant:** *Steven Ross* (University of Maryland)
12:30- 13:00	Writing on Admissions Tests and in University Classes: Two Constructs or One? *Brent Bridgeman.*	What is speaking made of? Comparing 9th graders' speaking performances in English and Swedish *Raili Hilden, Marita Härmälä*	
13:00- 13:30	Assessing oral proficiency in a foreign language *Natalia Ringblom*	Construct validation in a test of reading for young learners *Angela Hasselgreen, Torbjorn Torsheim*	
13:30–15:00	**LUNCH BREAK**		
15:00-16:30	**ANNUAL GENERAL MEETING** (Conference room: Ground floor-NEXUS- Building 6G)		
16:30- 17:00	**COFFEE BREAK**		
16:30- 17:00	**POSTER PRESENTATIONS** (Ground floor-NEXUS- Building 6G)		
	A framework for enhancing teachers' assessment literacy *Dina Tsagari, Karin Vogt, Ildikó Csépes, Tony Green, Nicos Sifakis*		
	Analysis of Reading Construct in PTE Academic *Abdullah Arslan, Salih Ozenici*		
	Motivation and attainment in English for tourism among university undergraduates *Zoltán Lukácsi*		
	Choice Given or Not?: Assessing Writing Skills of EFL Students *Meral Melek Unver.*		
	The interface of consequential validity and language assessment literacy *Dina Tsagari, Karin Vogt*		
	Investigating the construct of a writing test across languages: a corpus based approach *Michael Maurer, David Moreno*		
	Construct validation: challenges of a multi-level Language Program *Yevgeniya Pronoza*		
17:00- 18:00	**PARALLEL PAPER PRESENTATIONS** (2nd floor NEXUS- Building 6G)		
	Room 2.7 (2nd floor) Chair: *Slobodanka Dimova*	**Room 2.8 (2nd floor)** Chair: *Marisa Carrió*	**Room 2.9 (2nd floor)** Chair: *Norman Verhelst*
17:00-17:30	**WORK-IN-PROGRESS** Including Teacher Voices Through An Institutional Writing Criteria *Dilek Salki, Bahar Hasirci*	Classroom Assessment Construct: EFL Teachers' Perceptions and Practice *Hossein Farhady*	Tests of lexicogrammar: using expert judgements to investigate their underlying construct *Theresa Weiler*
17:30–18:00	An examination of how analytic raters adapt to holistic marking: a quantitative study *Judith Fairbairn*	Test-takers' voices in assessment tasks *Sehnaz Sahinkarakas*	Psychometric analysis: Evaluation of EAP Language Tests ítems and a Pre-sessional course *Ricardo de la Garza Cano*

20:30	**CONFERENCE DINNER & MUSIC - DANCE** (Hotel ASTORIA PALACE – Plaza Rodrigo Botet 5– *City Centre*)

SUNDAY 8th MAY

9:30- 11:00	**PAPER PRESENTATIONS** (Conference room : Ground floor - NEXUS- Building 6G) Chair*: John De Jong*
9:30- 10:00	Re-defining the construct of vocabulary size tests: Challenging Conventions *Benjamin Kremmel*
10:00- 10:30	Investigating the construct of speaking proficiency for young language learners: A discourse-analytic study. *Ching-Ni Hsieh, Yuan Wang.*
10:30- 11:00	What to test at C1? *Susan Sheehan*
11:00- 11:30	**COFFEE BREAK**
11:30-12:30	**ROUND TABLE & CONCLUDING REMARKS** (Conference room : Ground floor - NEXUS- Building 6G) **Topic:** *Summing up, thinking forward* **Moderator:** *Claudia Harsch* **Speakers:** *Neus Figueras, April Ginther, Constant Leung, Sauli Takala*
12:45	**CLOSING CEREMONY** (Conference room : Ground floor - NEXUS- Building 6G)
13:00	**TRIP TO *LA ALBUFERA LAKE***

PRE-CONFERENCE PARALLEL WORKSHOPS 3-5 May 2016
CONTENTS
Language Centre (CDL) Building 4P

TUESDAY 3: Registration 12:30-13:30 (CDL- Ground floor)

TUESDAY 3	WEDNESDAY 4	THURSDAY 5
13:30-17:30	10:30-13:30 15:00-17:30	9:30-12:30

WORKSHOP 1
Room Multimedia
(1st floor)

Online Testing Resources & Computer-based Assessment

Jenny Liontou: Greek Ministry of Education
Dina Tsagari : University of Cyprus, Cyprus

Overview

It is refreshing that our field is recognizing and beginning to discuss the importance and underlying theoretical and practical underpinnings of online testing resources and computer-based assessment, an area that is gradually coming into its own.

Objectives

The **objectives** of the proposed workshop are to provide information on computer-based assessment issues (terminological and procedural) and to give participants the opportunity to experiment with online instruments such as Mahara e-Portfolio System, Omnium classroom, Moodle, Riddle and Wikispaces along with open technologies such as Blogger and Twitter while reflecting on their use for assessment purposes.

Contents

The **contents** of the workshop will include**:**

a) basic information on aligning assessment with learning outcomes and designing assessment instruments for various learning purposes,

b) brief introduction to assessment of learners' language competencies, assessment of different language competencies and assessment in heterogeneous contexts,

c) considerations of using online resources to enhance classroom-based assessment while discussing its benefits to the instructor and the students,

d) issues concerning assessment feedback and practical affordances provided by open online technologies, and

e) using e-Portfolio as a reflective learning, teaching and assessment tool.

Target audience

The workshop is **targeted** at EALTA members (researchers, language testers, teachers and teacher trainers) who feel the need to have a better grasp of the most recent online testing tools and computer-based assessment principles and strategies and to share their experience with colleagues working at the same or other levels of education. Convenors of the workshop will employ a variety of **modes** to deliver the workshop such as mini-lectures, seminars that will involve participants in individual/group work and hands-on-practice with online tools that participants will be asked to use for scenario-based assessment purposes.

Background knowledge

Participants are expected to have no prior knowledge on computer-based assessment but they will be encouraged to contribute their expertise and teaching/testing experience to the sessions. Participants should have their laptop with them on Wednesday and Thursday.

Workshop facilitators

JJenny Liontou holds a Ph.D. in *English Linguistics* with specialization in Testing & Computational Linguistics from the Faculty of English Studies, National and Kapodistrian University of Athens and an M.Sc. in *Information Technology in Education* from Reading University, UK. **She also** holds a B.A. in *English Language & Literature*, a B.A. in Spanish Language & Literature and an M.A. in *Lexicography: Theory and Applications.* She has worked as an EFL tutor in online and distance-learning courses and as a freelance expert test consultant, item writer, oral examiner and script rater for various international examination boards. She has published widely and presented in numerous local and international conferences. Her current research interests include theoretical and practical issues of computational linguistics, on-line testing practices and classroom-based assessment. Jenny has experience in delivering workshops given the fact she has offered a range of workshops and seminars at the Faculty of English Studies, National and Kapodistrian University of Athens.

Dina Tsagari is an Assistant Professor in Applied Linguistics/TEFL with specialization in the area of Language Testing and Assessment (LTA) at the Department of English Studies, University of Cyprus, Cyprus. Dina is the director of the Language Testing and Assessment Lab of the University of Cyprus and the coordinator of the Classroom-based language assessment (CBLA) Special Interest Group – EALTA. Dina teaches undergraduate and postgraduate courses in language testing and assessment, EFL teaching methodology, qualitative research methods and supervises postgraduate students (MA and PhD) students in language testing and assessment at the Department of English Studies, University of Cyprus. Dina has experience in organizing and delivering training events. For instance she has successfully delivered another EALTA pre-conference, e.g. *Classroom-based assessment Pre-conference workshop (with Dr Neus Figueras Casanovas & Oscar Soler-Canela). 10th Annual Conference of the European Association of Language Testing and Assessment (EALTA), Istanbul, Turkey.* Dina has also taught on the *2nd EALTA Summer School* and on the *3rd EALTA Summer School, Università per Stranieri di Siena (Italy).*

WORKSHOP 2
Room 9
(Ground floor)

ASSESSING WRITING: Designing better rubrics, building better rating communities

Emma Bruce : City University of Hong Kong
Liz Hamp-Lyons : Editor of the journal Assessing Writing

Overview

As performance assessments play a greater role in language testing, especially in the contexts of classroom-based assessment and learning-oriented assessment, the role of rubrics for assessing writing and providing feedback grows too. This workshop aims to:

(1) provide participants with a clear understanding of what rubrics can and can't do, and the tools necessary for creating their own rubrics for their own contexts;

(2) introduce participants to effective processes for training raters to assess writing based on rubrics designed for their own context;

(3) demonstrate qualitative and simple quantitative quality assurance procedures evaluating the validity of the rubric and the reliability of the ratings.

Intended learning outcomes

By the end of the workshop, participants will be able to:

- Understand the difference between a rubric and a rating scale;
- Identify different types of rubrics and their applications for different purposes in assessing writing;
- Critique a range of rubrics, given contextual information;
- Understand the essential factors involved in rubric choices;
- Design a contextually-appropriate rubric;
- Apply simple quality assurance techniques to evaluate the rubric;
- Understand the importance of data-driven rubric and scales;
- Implement effective rater training procedures in their own contexts;
- Apply simple quality assurance measures to ensure the effectiveness of on-going rater training.

Contents and methods

Discussion topics

- Rationale for performance assessments
- Components and terminology of a performance assessment
- Rating scales and rubrics for assessing writing
- Validity and reliability – quality assurance
- Fairness in design and in score uses

Activities

- Critiquing rubrics for purpose and context
- Applying procedures for selecting or developing scales / rubrics;
- Participating in the creation of a rubric for a shared context/ purpose;
- Working with others to create a rubric /scale, and critique it;
- Reading, discussing, ranking and scoring written performances
- Matching exemplars to descriptors
- Participating in and leading mock rater training sessions
- Analyzing rater statistics

Background knowledge/pre-workshop activities

Participants should have the required background and prior knowledge:

- Experience in teaching English as a foreign/second language, or language development coursework
- Strong interest in using performance assessments for judging and reporting written performance
- Some familiarity/experience with performance-based assessment of writing or speaking
- Participants are encouraged to bring at least one example of a rubric/rating scale they have used, and 4-10 sample written and scored performance responding to a writing task within their own context.

Pre-workshop reading

Bruce, E. & L. Hamp-Lyons. (2015). Opposing tensions of local and international standards for EAP writing programmes: Who are we assessing for? *Journal of English for Academic Purposes, 18,* 64-77.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 162-189). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (2007). Worrying about rating. (Editorial). *Assessing Writing, 12,* 1-9.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing 16,* 81-96.

Weigle, S. C. (2012). Assessment of writing. *The Encyclopaedia of Applied Linguistics*. Pub. Wiley. Available from http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0056/full

Workshop facilitators:

Emma Bruce runs the Assessment programme at the English Language Centre of City University of Hong Kong. In this role she oversees the development and implementation of new assessments, the quality assurance of on-going assessment practices, and the day-to-day operations involved in the delivery of fair assessments and exams in a large university language centre. Emma's main research interest is in integrated writing assessment for EAP and she is currently pursuing her PhD studies.

Liz Hamp-Lyons is the Editor of the journal *Assessing Writing*, and edited the first–ever collection of research in second language writing assessment, titled *Assessing Second Language Writing in Academic Contexts,* which included five chapters based on her PhD. She has been researching writing assessment, working to develop and validate writing assessments, and training raters to assess writing since 1984.

WORKSHOP 3
Seminar
(1st floor)

Quality assurance in test development: a hands-on introduction to task specifications, item writing, trialling and statistical analysis

Gwen Caudwell: Aptis Product Development Manager, British Council
Kevin Rutherford: Aptis Test Production Manager, British Council
Judith Fairbairn: Test Development Consultant, British Council
John Tucker: Aptis Test Production Co-ordinator

Intended learning outcomes:
This practical workshop will help participants learn some of the important quality assurance techniques that professional test developers employ when designing tasks and specifications, writing items, trialling tests, and conducting statistical analysis.

Content:
A hands-on workshop where the Aptis team will take participants through the steps to create one reading test.

- Part 1 (3 May): Task specifications
- Part 2 (4 May): Item writing
- Part 3 (4 May): Trialling
- Part 4 (5 May): Statistical analysis

Methods:

- An expert in each area from the Aptis team will deliver each part (we will pay for the extra people to attend).
- Participants will be placed into five groups with six people in each group.
- Aptis presenters will circulate and sit with groups, assisting where needed.
- All activities will be done on-line using SurveyMonkey and Google docs (or similar online technology).
- Part 1: Designing and developing a reading task and task specifications.
  - Brief introduction to designing and developing a task and task specifications. Participants will be given a handout on how to write good specifications and a specifications template.
  - Each group will develop one task and write specifications for the task. The task must be no longer than 15 minutes to answer.
- Part 2: Item writing.
  - Brief introduction to item writing. Participants will be given a handout on how to write good quality items and use on-line analysis tools.
  - The task specifications created in Part 1 are passed to a different group.
  - Groups will write items in SurveyMonkey (or similar) using the task specifications.
  - Groups will share their items with the group that wrote the specifications and discuss any issues.
- Part 3: Trialling a task
  - Brief introduction to trialling.
  - Groups take the three tasks they have not worked on, responding to the task items. The full test will have been sent the night before to a pre-selected group of test takers in order to collect enough data (or dummy data will be created for the statistical analysis).
  - All marks are automatically collated into a shared template for analysis in Part 4.
  - Discussion to include any problems with items and tasks.
- Part 4: Statistical analysis techniques for trial results.
  - Brief introduction to statistical analysis of receptive skills.
  - Steps to analyse data
    - Descriptive statistics: participants will use excel to calculate descriptive statistics and histograms. Step-by-step handout provided and participants can practice steps.

- How to create item labels, person labels and a concurrent worksheet in excel for Winsteps. Step-by-step handout provided and participants can practice steps.
- Create a Winsteps Control File, Item Map and Item Measures. Demonstration only.
- Discussion to include issues with the data analysis and the test created by the group.

Background knowledge/pre-workshop activities

No specific knowledge required but participants must have a laptop with internet and be setup with Google Docs for sharing work.

Pre-workshop activities: None required.

Workshop facilitators:

Judith Fairbairn: Test Development Consultant, British Council

She has an extensive ten years of testing experience in managing examiners for IELTS and Aptis. Her areas of expertise include examiner management global policy creation and marking quality assurance, and she holds a certificate in Theory and Practice in Language Testing from Roehampton University.

She is currently studying an MA in Language Testing focusing on examiner marking quality assurance and creating on-line examiner support, training and standardisation systems. Judith was invited to the International Association of Applied Linguistics 2014 conference to deliver a presentation on Quality Assurance in online, international Large Scale Testing. In addition to this, she is also an experienced examiner, item writer, and has provided editorial assistance for IELTS Research Reports Volume 12.

Gwen Caudwell: Aptis Product Development Manager, British Council

Ms. Gwendydd Caudwell is Aptis Product Development Manager for the British Council based in Dubai. In 2009 she took the Roehampton Theory and Practice of Language Testing course as a means to understand the rationale behind some of the tests she was using with her students in her teaching career.

This sparked her interest in this area and she completed her MA in Language Testing with Lancaster University in 2012. Gwendydd has extensive experience in teaching, teacher training and management for the British Council around the world.

She is responsible for the research and design of new variants of Aptis and British Council tests and has also worked on designing and implementing tests in different contexts such as the Science Olympiad Foundation in India and a Young Learner test for a large project in Uruguay. Her particular areas of interest are rater judgements and Young Learner assessment and is looking to base her PhD on this topic.

Kevin Rutherford: Aptis Test Production Manager, British Council

Kevin Rutherford is currently the Aptis Test Production Manager, and is based in Warsaw, Poland. He is responsible for managing the production of content for Aptis tests and for the trialling of test items, and the setting up of Aptis test centres.

He also leads on the training of item writers and is involved in the development of new test items. Before moving into the field of assessment, he had a long career in English language teaching, in Japan, Italy, and the United Kingdom, and, since 1996, in Poland.

He completed the Certificate in the Theory and Practice of Language Testing with the University of Roehampton in 2010, and was subsequently involved as an item writer in the early development of Aptis. He obtained an MA in Language Testing from the University of Lancaster in 2014. He has given presentations on assessment at conferences for the International Association of Teachers of English as a Foreign Language (IATEFL) and the European Association for Language Testing and Assessment (EALTA).

In 2014 he co-presented a workshop on item writing and test specifications at the Language Testing Research Colloquium (LTRC).

John Tucker has been Aptis Test Production Co-ordinator, based in Poland, since July 2012. His current responsibilities include commissioning and quality reviewing test items for Aptis General and other tests, along with helping to produce the tests themselves. John is involved in

producing materials for and training item writers, and presented at ALTE 2014 in Paris on this subject. He also handles technical queries connected with the administration of Aptis tests. John is part of the British Council's new Assessment Literacy Project, for which he is writing a module on listening; John also co-presented on this topic at a pre-conference TEASIG workshop in Harrogate. Earlier in 2014 he took an open access course in corpus linguistics run by Lancaster University.
John completed an MA (distance) in Language Testing as well as the BC's E-moderator Essentials course in December 2013. Before that he took a diploma in testing from Roehampton University. Previously, John ran placement testing workshops and helped re-design assessments for the British Council's Korea centre in Seoul. Prior to that, John had a long and varied career as a teacher in the UK, Colombia, Spain and Thailand.

WORKSHOP 4
American Space
(1 floor)

A practical approach to tackling challenges in Quality Assurance (QA) in test development projects

Neus Figueras: Generalitat de Cataluña
Elaine Boyd: Institute of Education, University College, London

Overview
The workshop will focus on the challenges that are embedded in the test development process and explore a variety of approaches and possible solutions. This workshop is for those who are involved in test development at whatever level.
Issues of quality are what can make or break a test and the impact can be far reaching as more and more 'accountabilities' are attached to tests. Yet the scale and scope of quality processes can seem daunting to many language certificate awarding institutions, especially small scale operations. The objectives of the workshop are to analyse the many facets of quality and to explore challenges in both implementing and sustaining quality. The main aim, however, is to provide a discussion forum to share and learn, and to help build well-grounded confidence in QA procedures.

Learning outcomes:
- To agree on what quality means in different contexts.
- To plan for effective consultation and review in QA.
- To manage and prioritise the contributions of both stakeholders and experts.
- To understand the relationships between data and expert opinion.
- To develop strategies for tackling challenges and risks to QA.
- To identify possible threats to sustainability.

Contents
Session 1: CONSULTATION
- What does QA look like and why does it matter?
- How can we plan for quality (even if a small scale test)?
- What do different parties want to see? (eg regulators / Ministries / Govt offices / Executive)
- Who do you need to 'recruit'? What is the profile of an Expert?
- Why is consultation important?
- How can belonging to professional networks help?

Session 2: PROCESS
- Where does your test sit? How do you know?
- What are the key steps in QA?
- 'What if …?' scenarios. How to plan ahead?
- What can't you know/control?

Session 3: REVIEW
- What is the difference in QA in testing receptive vs productive skills?
- Examiner standardisation - what can go wrong? How to fix?
- What claims can you make?

Session 4: CHALLENGES
- What challenges do different contributing groups (eg senior examiners, external experts, etc) can bring to the process?

- What is the danger of making assumptions (eg using 'known' items; introducing a new test to experienced examiners, etc.)?
- How do we deal with special needs and complaints?

Approach

The sessions will focus on the sharing of best practice(s) through the discussion of concrete examples presented by the workshop leaders. Participants will also have an opportunity to present their own issues, queries or problems to the group for discussion and potential resolution. Using the approach captured in the "Mantle of the Expert" (Dorothy Heathcote, 1985), participants will analyse a series of problems and situations to discuss how to manage and improve QA. The approach aims to replicate the stages of consultation, critical feedback and review and so reflect a 'loop input' of what a QA process might look like.

Target audience

EALTA members and interested individuals who are:

- involved in the development of an assessment system
- interested in improving QA/standards of training.

Background knowledge

No background in Test Development is assumed although it is anticipated participants have an engaged with design or development of tests at whatever level.

Pre-workshop preparation

Participants are encouraged to bring case studies or challenges to discuss with the group. Cases /challenges can be anonymised as we aim to extract the general from the particular. These do not have to be from a full-blown test development and may simply revolve around issues working with a very small team, limited resources etc.

Workshop leaders will communicate with the participants before the workshop and send documentation/readings which may be useful to consult in advance

Workshop facilitators:

Neus Figueras has worked in the Departament d'Ensenyament de la Generalitat de Catalunya coordinating the certificate exams for the EOI for 20 years. She lectures part-time at the University of Barcelona and the Universitat Pompeu Fabra. She has been involved in a number of international research and development projects (Speakeasy, Dialang, Ceftrain) and collaborates regularly with the Council of Europe in the dissemination of the Common European Framework of Reference in relation with testing and assessment. She has published articles in the field of language teaching and assessment and is one of the authors of the Manual for Relating examinations to the CEFR (Council of Europe, 2009). She has recently published, with Fuensanta Puig, Pautas para la evaluación del español como lengua extranjera (2013). Edinumen. She has been a teacher trainer for over 20 years, and has given courses and presented in universities in Spain and in different European countries, in Asia and the USA. She was the first President (2004-7) of EALTA (European Association for Language Testing and Assessment), and she is now an expert member (www.ealta.eu.org)

Elaine Boyd has worked in assessment design and development, examiner training and standardisation and quality standards for over 20 years for a range of international testing organisations, including Cambridge English and Trinity College London. She has worked on exams across a range of levels and domains and has published several exam course books and test practice books. She has also conducted courses and sessions in item writing and assessment literacy for teachers in Europe and India and has published articles in this field. She holds a PhD in spoken language and pragmatics from the University of Cardiff and is an Associate Tutor for the online MA in Applied Linguistics and TESOL at Leicester University. She is currently working on the Trinity Lancaster Corpus of Spoken Language.

SIG MEETINGS 5 May 2016

THURSDAY 5 14:00- 17:30

Language Centre (CDL) Building 4P

SIG 1 Room 3 (Ground floor)	Assessing Speaking
SIG 2 Room 9 (Ground floor)	CEFR
SIG 3 Seminar(1st floor)	Special Interest Group 'Classroom-based Language Assessment'
SIG 4 American Space(1 floor)	Assessment of Writing and Assessment for Academic Purposes

EALTA's 13th CONFERENCE- 5-8 May 2016

Assessment of what...? Revisiting the issue of construct(s)

CONFERENCE PROGRAMME: ABSTRACTS

THURSDAY 5th MAY

12:00-18:00	**REGISTRATION** *(Ground floor CDL – Building 4P)*
18:00-20:00	**WELCOME RECEPTION** at THE LANGUAGE CENTRE *(Ground floor CDL – Building 4P)*

FRIDAY 6th MAY

8:30-9:45	**REGISTRATION** *(Ground floor NEXUS – Building 6G)*
9:45- 10:00	**OPENING CEREMONY** **Room:** Conference room *(Ground floor NEXUS – Building 6G)*
10:00-11:00	**KEY NOTE PRESENTATION** **Room:** Conference room *(Ground floor NEXUS – Building 6G)* Chair*: Cristina Perez-Guillot*
10:00-11:00	**Constructing Language and Literacy Practices– Implications for Assessment** ***Constant Leung:*** *Professor of Educational Linguistics*

The construct of language proficiency in our time has been largely understood in terms of lexico-grammatical knowledge and the ability of an individual to make use of it in communication. At the same time, communication, or more precisely communicative acts such as discussion in a meeting, has been seen as comprising a set of typical characteristics that is routinely enacted, irrespective of context and participant volition. While this approach has been very helpful in bringing together structural and socio-pragmatic aspects of language use, there is growing recognition that these typifications can only provide a partial account of the dynamic and contingent nature of language use in communication. *In extremis*, this approach can reify language proficiency as a set of generalizable scripted productions. In this talk I will exemplify this by looking at relevant research in academic language and literacy in higher education.

Recent research in Academic Literacy/ies has consistently shown that there is considerable diversity in language and literacy practices across different disciplines. Empirical accounts of the lack of correspondence between the tasks in some academic language tests and the actual in-course language use point to the problems of over-relying on generalized *pre*-scription (e.g. Elder, 2007; Paul, 2007). In the first part of my talk I will examine the construct of academic language and literacy as it has been represented in large-scale assessment frameworks (e.g. CEFR) against the backdrop of some of the empirical findings from the field of Academic Literacy/ies with particular reference to higher education (e.g. Lea and Street, 1998, 2006; Tribble and Wingate, 2013; Wingate, 2015). I will then argue for the need to recognise multiple constructs of academic language and literacy and to take account of the diverse processes of *doing* academic literacy. In

the last part of the talk I will discuss the implications of a situated, practice- and process-informed view of construct for the assessment of academic language and literacy

11:00–11:30 COFFEE BREAK

11:30- 13:30 PAPER PRESENTATIONS

Room Conference room *(Ground floor NEXUS – Building 6G)*

Chair*: Neus Figueras*

11:30- 12:00 Language assessment literacy for test developers: towards an understanding of test constructs

Cristina Rodriguez, Julia Zabala

Construct validation is a well-researched topic in language testing, particularly in the case of high stakes tests in which the presence of construct irrelevant factors can have serious consequences for the candidates involved. These studies frequently involve correlations between theoretical models of language ability and test scores and identify construct irrelevant items or subsets. However, less attention has been traditionally paid to the role played by test developers in establishing solid and consistent constructs, and specifically, to the LAL of test developers with regards to the "principles and concepts that guide and underpin practice" (Fulcher, 2012).
Our paper presents a comparative study carried out in Spain amongst test developers in two of the largest institutions developing language tests at a national level for adults: EOI and University Language Centres. Our goal was to analyse the level of LAL of test developers based on the amount and characteristics of their training, the impact of this training on test developers' work, and their own perceived training needs. The results of this study can shed light on the understanding of test constructs and be used for the preparation of training programmes that aim to support the development of LAL that goes beyond the acquisition of test-writing skills to include a solid understanding of the principles that underlie the test.

12:00- 12:30 "Study on comparability of language testing in Europe": focus on the constructs

Esther G. Eugenio, Michael Corrigan

The results of different language tests are sometimes treated as equivalent and an assumption is made about the comparability of the constructs measured. However, such assumptions are not always well-founded and merit further examination. This presentation reports on a study completed by Cambridge English Language Assessment for the European Commission to investigate the comparability of the results of 133 national language tests across Europe, and more specifically on the way in which the assumption of comparability of constructs was examined.
Following the "Conclusions on Multilingualism and the Development of Language Competences" (Council of the European Union, 2014) adopted in May 2014, the European Commission requested a study to compare 133 national language examinations at levels ISCED 2 and ISCED 3 (lower and higher secondary education) from 33 independent educational jurisdictions (28 EU Member States). The languages included were EU official languages studied by more than 10% of the students in each jurisdiction. After collecting all exam materials and supporting documents, a group of trained experts in language assessment from across Europe analysed each exam using an online content analysis tool. This data was then analysed statistically to appraise test comparability.
The findings from this study (Cambridge English Language Assessment, 2015) show considerable diversity in the constructs tested across all the different national language examinations, despite common terms being used to name test components (e.g. 'Reading'). It is therefore misleading to assume that results from national language examinations can be compared and used interchangeably since they are actually testing different constructs.

12:30- 13:00 **An empirical investigation of the relationship between L2 vocabulary size and L2 listening comprehension at the B1 CEFR-level in English and French.**

Britta Kestemont, Ann-Sophie Noreillie, Kris Heylen, Piet Desmet, Elke Peters (*Joint First author)*

Vocabulary knowledge is a key predictor of language proficiency (Schmitt, 2008). However, little research has looked into the relationship between vocabulary size and listening comprehension. One exception is Staehr's study (2009) which showed a strong relationship between vocabulary size and listening performance among advanced English-as-a-foreign-language learners (C2 CEFR-level). However, it is not clear whether this would also be the case for other proficiency levels and other foreign languages. Therefore, this study aims to investigate the relationship between L2 vocabulary size and listening comprehension at the B1-level in two foreign languages, English and French.
In this study, 199 English-as-a-foreign-language learners and 351 French-as-a-foreign-language learners, recruited from secondary schools and first-year university students, took part. Vocabulary size was measured by means of a frequency-based multiple choice test for English and French. Listening comprehension was tested in the PET (B1; Cambridge) for English and the DELF (B1; CIEP) for French. Our results indicated a strong correlation between vocabulary size and listening comprehension in both languages. Moreover, a preliminary analysis of the lexical profile of the English and French listening tests revealed that knowledge of the first 1,000 words corresponds to 95% (English) and 86% (French) lexical coverage. By comparing two languages and different proficiency levels, our results may refine our understanding of the role of vocabulary size for listening.

13:00- 13:30 **French speakers reading German and German speakers reading French – do they take different tests?**

Monique Reichert, Charlotte Krämer

The past two decades have seen an important number of studies evaluating reading comprehension. However, it has become increasingly difficult to compare results from different studies, deploying different tests, targeting different languages and persons with different language backgrounds. A better knowledge of the attributes that are responsible for the difficulty of reading comprehension items would hence lead to a more theory-driven test specification. The current study aims at exploring the benefits of using the linear logistic test model (LLTM; Fischer, 1973) regarding the identification of those cognitive, linguistic and test-specific attributes that best describe and explain reading test performance. Strongly based on scientific findings from reading literacy studies, a list of attributes (e.g., making inferences, position of correct option in MC item) was determined. The same list was used to specify both 33 German and 34 French reading comprehension items with known and adequate statistical parameters. This specification resulted in a Q-matrix, in which each attribute was given a weight, depending on whether the attribute was essential or not for solving the item. Finally, the LLTM was applied to the data of Luxembourgish/German, French, and Portuguese speaking 9th graders that had previously taken the German and the French reading test. The results from the LLTM modeling show a significant overlap between both tests, and between the three language groups regarding the strength of the attributes in explaining item difficulty. The findings help explain the construct of reading comprehension and are discussed regarding their implications on test construction and teaching.

13:30–15:00 LUNCH BREAK

15:00- 16:15 WORK-IN-PROGRESS SESSIONS

Room 2.7 (2nd floor)

Chair*: Jamie Dunlea*

15:00- 15:25 Editing tasks – (what) do they add to the writing construct?

Sonja Zimmerman, Anika Müller-Karabil

Integrated writing tasks are already common practice in large-scale assessments and have been researched from many perspectives (Cumming 2014; Plakans 2012). However, the nature of the underlying construct is still an open issue. Advances in the field have been made by considering 'typical' reading-into-writing tasks where examinees have to prove their ability to write from one or more language-rich source texts. Findings have indicated that integrated writing tasks measure writing ability rather than reading and/or listening (e.g. Asención Delaney 2008; Gebril 2010; Sawaki, Stricker & Oranje 2009). But does this also apply to tasks that only require a limited written output for completion, e.g. writing one-sentence summaries or editing a paragraph?
This work-in-progress presents preliminary results of an exploratory study into the construct underlying editing tasks by addressing the following research questions:

(1) To what extent do the results of editing a paragraph correlate with writing ability? That is, can editing tasks contribute to representing the construct of writing ability, and can they be regarded as "a useful complement to free-response writing tasks" (Breland et al. 2001)? Or is the underlying ability of an editing task more related to reading ability?

(2) Does the underlying ability change in relation to the kind of errors that have to be corrected? For example, are high-order concerns (i.e. global revision) more related to writing ability than low-order concerns (i.e. local revision)?

Further steps for investigation into the construct of integrated limited production tasks will be offered for discussion.

15:25- 15:50 The construct of writing across contexts.

Carol Spöttl, Sonja Zimmerman, Jayanti Banerjee

For assessment purposes, the ability to be tested needs to be clearly specified. The definition of the construct is therefore a fundamental consideration in the test development process. The challenge of L2 writing assessment is to identify skills and key features that represent writing in all its complexity, so that they can be applied across a wide range of writing situations and communicative purposes, without neglecting the context-specific aspects of different kinds of writing tasks.
This work-in-progress focuses on the assessment of L2 writing proficiency at the level B2 of the CEFR by addressing the following research question: How and to what extent does the context of a particular writing assessment affect the operationalization of the identified key aspects of writing ability at this specific level?
Preliminary results from a qualitative analysis of standardized writing assessments in two different contexts are presented:

- assessing L2 writing proficiency as a language requirement for university admission for international students; and,
- assessing L2 writing proficiency as part of a school leaving exam.

In context one the exam is not linked to any specific curriculum the language tested is the medium of instruction in the target university settings. In context two, however, the exam assesses L2 writing proficiency at the end of an educational programme and the language acts as proof of foreign language ability, not necessarily being the medium of instruction.
Finally, a preliminary research agenda for capturing the key features of the B2 writing construct will be offered for discussion.

Room 2.8 (2nd floor)

Chair*: Dina Tsagari*

15:00- 15:25 ECD for MSA: Developing comprehensive construct definition

Bjorn Lennart Norrbom, Abdulrahman al-Shamrani, Yong Lou

The need to develop a comprehensive construct definition for Modern Standard Arabic (MSA) stems from current test practices as well as from the field at large as MSA constructs are typically underspecified for both teaching and testing causing test developers to rely heavily on psychometric evidence for validation purposes. MSA construct definition is complicated by a strong element of diglossia; particularly for listening and speaking where various colloquial forms, not MSA, dominate most Target Language Use (TLU) domains. The paper describes the validation of a test covering listening, reading, writing, and grammar with the aim of sufficiently describing the test construct through domain analysis and modelling using Evidence Centered Design (ECD). At the core of ECD is the Conceptual Assessment Framework (CAF) where the most relevant parts in relation to domain modelling are the Student Model (SM) (*what* we measure), the Evidence Model (EM) (*how* we measure), and the Task Model (TM) (*where* we measure). For the CAF to support test validation, it must rest on a foundation of domain modelling typically lacking for MSA. This prevents drawing valid inferences between the TM and the SM. Consequently, psychometric evidence cited in the EM also comes into questioning augmenting the need for thorough domain modelling. Research started with careful domain analysis and modelling using a variety of sources and then moved on to start creating the CAF. The paper carries the potential to stimulate discussion about construct definitions in Arabic and other less frequently tested languages and diglossia in language testing.

15:25- 15:50 25 years of assessment at the EOI (Government-owned Language School) in Malaga.

Carmen Medina.

The government owned language school in Málaga started in 1970. Since that year and until 2012, when the Andalusian educational authority started sending unified exams for the whole of the region, it has been creating its own exams via the team work of the teachers who were posted there during each academic year. This paper revises the evolution of the exams from 1986 to 2012; it compares the initial levels of the exams with the current ones. It also goes over the structure, contents and exercise types. It looks at what is assessed and how it is assessed. It then goes over the standardized sample models supplied by the Andalusian educational authority from 2012 to June 2015 to see what is assessed and how it is assessed before comparing the shifts in structure, contents and exercise types.

15:50–16:15 Towards Localisation of a Test: What are the Essential Considerations?

Ying Zheng, Yanyan Zhang

This study investigated potential Aptis test takers' perceptions of its validity and test practicality in China. Aptis test is a computer-based test developed by British Council, assessing grammar, vocabulary and four language skills. 84 students from Wuhan University China participated in the study. Mixed methods approaches were adopted. A questionnaire was used to investigate test takers' perceptions on the aspects of their test-taking experience, test items, skills measured, and some practicality issues. Semi-structured interviews were carried out to further probe into participants' perceptions of the test.

This study was designed to evaluate whether contextual, social and topical dimensions exert any influence on impacting the validity of this test, as well as to understand how English skills (constructs) measured in Aptis can be compared to the skills measured in other English tests in China. In particular, the study looked into test takers' performance comparing their gender and university major differences; their experiences in taking Aptis in terms of the perceived difficulty levels and opinions of taking a computer-based test of this nature; and their perceptions of Aptis and what the perceived differences are from those of other national or international tests that this group of test takers had experienced before.

Gaining a better understanding of test takers' perceptions of this new test can provide valuable information in informing test development as well as test preparation practices. Taking essential considerations into account, it is step towards the evaluation of the need for localisation of a test.

Room 2.9 (2nd floor)
Chair*: Marta Genísi*

15:00- 15:25 Developing an adaptive diagnostic test using evidence-centered design

Evelyn Reichard, Ingrid Bresser, Wilma Vrijs

Our Ministry of Education, Culture and Science has commissioned the development of an adaptive and diagnostic test for students who are halfway through their secondary education. This computer-based, formative test is aimed at diagnosing a student's skill level with regard to writing and reading in Dutch and English. The results provide insight into the students' strengths and weaknesses on various domains in order to help them improve their performance. The student, teacher and school each receive their own report. The test is based on the theory of evidence-centered design by Mislevy, R. J. and others, which provides a solid framework for construction and reporting on the results. The adaptive model is based on Bayesian statistics, which makes it possible to report on the students' skill levels. The test gives a reliable prediction of students performing below, at or above their expected skill level. Our presentation focuses on the development of a diagnostic test using evidence-centered design for construction. The combination of formative, diagnostic and adaptive elements is new and the development process has been very instructive. We will also discuss how Mislevy's theoretical framework was used as a basis and how it was combined with expert knowledge from the field to develop a valid content design. We will show some of the question types that were developed specifically for the test. And we will explain how the use of adaptivity makes it possible to give accurate predictions of skill levels in an efficient way.

15:25- 15:50 Assessment of what ...? Validating a large-scale, high-stakes speaking test

Thomas Koidl

This paper presents some of the findings of an empirical evaluation study concerning the speaking component of the reformed school leaving exam. The reform introduced standardised, centrally developed written exams and competence-oriented oral exams for the first time in 2015. In modern languages teachers now administer a self-designed proficiency test with an individual long turn (ILT) and a paired activity (PA) instead of the previous achievement test.
It seemed worthwile investigating
(1) how uniformly (or not) the test was administered
(2) how teachers handled their new role of interlocutor
(3) how they applied the new assessment criteria.
Therefore, the study addresses issues of construct validation.
A questionnaire was designed covering these three aspects of the exam. Closed response and Likert scale like questions were used with an open-ended question for comments at the end. The respondents were teachers of English, French, Italian, Russian and Spanish involved in the oral exams either as interlocutors or assessors or both. Some 950 questionnaires were returned through the official channels they were sent out. All provinces are represented in the sample.

Quantitative analyses using SPSS are currently carried out as well as qualitative interpretation of the open-ended responses (comments). In select cases it was possible to audio-record live exams the analysis of which, though not being part of this study, will hopefully shed more light on the validity of the exam. There are some serious concerns regarding the CEFR level claims we are making, especially in the first foreign language, English.

15:50–16:15 Social science reading constructs and the CEFR reading level of textbooks

Barbara Blair, Kari Telstad Sundet

The Norwegian compulsory education reform of 2006 specified five basic skills for learning in school, work and social life: oral, reading, writing and digital skills, plus numeracy. These skills are considered fundamental to learning in all subjects, as well as a prerequisite if pupils are to

demonstrate their subject competence. Thus, language skills are required in order to learn, for instance, history. By implication, history teachers are required to facilitate their pupils' development of general language skills and acquisition of subject-related academic language.
The aim of this presentation is to compare the reading constructs described in the social science curriculum goals for the 4th, 7th and 10 grades, with the CEFR reading level necessary for reading textbooks in this subject.
Six social science textbooks have been analysed. These include two textbooks for the 4th grade, two for the 6th grade and two for the10th grade. Six raters who are well acquainted with the CEFR have assessed these books to identify the CEFR level required in order to read and learn from them. A rating scheme based on the Dutch CEFR Construct Project (Alderson, J. C. et al, 2006) was used. The language employed in the textbooks was analysed with respect to such key criteria as vocabulary, grammatical complexity, language functions and text length.

The initial findings indicate that the CEFR level in reading competence that is required to cope with the language in the textbooks studied corresponds with the reading constructs specified in the social science curriculum goals.

16:15–16:45 COFFEE BREAK

16:45-17:45 PARALLEL PAPER PRESENTATIONS

Room 2.7 (2nd floor)
Chair*: Carol Spöttl*

16:45- 17:15 **Factors affecting listening: operationalising a listening construct**

Julia Zabala, Cristina Perez-Guillot

Listening is a complex process as its operationalization integrates both physiological and cognitive functions (Field, 2002; Rost, 2002; Wolvin, 2010). The reception of the message is not only affected by the listener's working memory, but also by his/hers perceptual filter –background, experience, mental and physical states-, turning listening into a skill that can be examined from multiple perspectives: neurolinguistics, cognitive psychology, language pedagogy, etc. The purpose of this paper is to analyse the operationalization of the listening construct by taking into consideration the factors that affect the listening comprehension process.

The study of the listening skills has been traditionally set aside since listening in our first language (L1) requires little effort and is acquired at infancy and there is a tendency to believe that this is the same for a second language. However, by treating listening in the L2 the same as in the L1, the additional processes that have to be performed by L2 listeners to overcome comprehension barriers are being disregarded. The objective of this study is to examine the study of such factors in the literature and present a practical application to the improvement of listening test tasks.

17:15–17:45 **Defining the listening construct in multimodal environments**

MariCarmen Campoy-Cubillo, Mercedes Querol-Julián

The *International Listening Association* defines listening as "the active process of receiving, constructing meaning from, and responding to spoken and/or non-verbal messages" (ILA, 1995:4). Drawning on this understanding of the listening process, we set forth the definition of a listening construct that considers different communicative modes in a mixed skills approach. This listening construct takes into account available listening material sources and formats, and possible response modes. We need to take into account that listening situations respond to either an interactive or a transactional nature depending on whether a (more or less immediate) response on the part of the listener is expected or not.

Input sources for listening tasks can be audio, video and face-to-face communicative situations. Traditionally, listening skills are assessed by means of using audio recordings where attention is paid to specific and general information conveyed through verbal content. The use of video as listening input (Campoy-Cubillo and Querol-Julián, 2015), though present in classroom tasks is less common in a language testing environment, includes non-verbal cues (visual, kinesic, contextual, etc.) as part of the components of the listening construct and gives learners the opportunity to face significant real life situations. Finally, in face-to-face communication we can assess the listener's response or reaction to the speaker's turn.

We propose a multimodal listening construct definition, based on video as input, that includes: (a) the examination and description of different input and output modes and its inter-relation, and (b) a multimodal question typology and question sequencing options.

Room 2.8 (2nd floor)
Chair*: Asunción Jaime*

16:45- 17:15 **Flip teaching and construct irrelevant variance in classroom based assessment**

María Luisa Carrió Pastor

Construct irrelevant variance (CIV) is the introduction of abnormal, uncontrolled variables that affect assessment outcomes. In the case of flip teaching, students work at home contents that are assessed later during class time. The aim of this study was to investigate the extent to which flip teaching can affect the outcomes of students in classroom based assessment. The CIV was investigated to avoid distortion when assessing flip teaching activities. The activities carried out by 60 students enrolled in a subject of English for specific purposes at Universitat Politècnica de València were assessed during the first semester of the academic year 2015-16. In this assessment procedure, two English teachers judge the different activities carried out by students and the amount of variation of the different activities was calculated. It was analysed if assessors assigned the same score to all students' performance or if more variation was found in the assessment of flip teaching activities. A considerable amount of variation was found in the evidence reported by both teachers. Classroom based assessment was not carried out equally when flip teaching was incorporated in the subject assessment. In this sense, it was concluded, after the analysis of the results, that the scoring guide and the conceptual framework for assessing teachers' coaching competence should be revised when new teaching techniques are incorporated in classroom based assessment

17:15–17:45 **Using a corpus to validate test construct: The case of epistemic markers in the Trinity Lancaster Corpus**

Elaine Boyd, Vaclav Brezina

By allowing us to study much larger data sets of language – and in more complex ways – corpora can be used to support test validation. This study focuses on The Advanced subcorpus of the Trinity Lancaster Corpus of L2 speech constructed from the Graded Exam in Spoken English (GESE) developed by Trinity College London. The corpus is annotated for both the part of speech as well as for a range of speaker variables such as L1, age, gender, education background, etc.
At the Advanced levels, the GESE includes a prepared presentation task and discussion, a collaborative task where the test taker takes responsibility for the interaction and a conversation where the test taker's views are challenged. The study explores validation through the use of epistemic markers across the four tasks. Epistemic stance (expressions of certainty and uncertainty) is an essential part of natural communication; it can show how successful learners are in natural discourse interaction and meaning negotiation (cf. Kärkkäinen 1992; Aijmer 2002). From the language testing perspective, this is important because we can assess how far the construct behind each task is operating as intended and eliciting different types of language.

The findings show that the production of epistemic markers differs significantly according to the type of task. The results not only help validate the claim that each task within the GESE Advanced

levels targets a different communicative construct but also illustrate how test takers adjust their speaking style when faced with different communicative requirements.

Room 2.9 (2nd floor)
Chair*: Carolyn Westbrook*

16:45- 17:15 **Defining vocabulary as a unitary construct: a CEFR-based framework**

Veronica Benigno, John De Jong

Although research on vocabulary acquisition (e.g. Read, 2000; Schmitt, 2000; Nation, 2001; Meara, 2005; Laufer, 2009) has evolved substantially over the last decades, there is little agreement on sequencing vocabulary teaching, i.e., the selection of words to be taught at increasing proficiency levels. The question therefore remains how to build vocabulary and measure it as a unitary construct.
This paper reports on the development of a graded lexical inventory developed using frequency analysis of L1 written and spoken data and teacher ratings. In a first step, the frequency of occurrence of 40,000 word meanings was retrieved from a 2.5 billion words reference corpus including written and spoken data. In a second step, word meanings were semantically annotated using the Council of Europe Vantage Specifications' categorization in Specific Notions, General Notions, and Functions (Council of Europe, 2001). Each word meaning was then rated by 19 teachers using a scale of 1 to 5 ranking communicative usefulness of vocabulary. Eventually, frequency data and teacher ratings were combined in a weighted model to scale vocabulary on the CEFR.

In response to the Council of Europe recommendation to the state members to create inventories of linguistic forms (known as Reference Level Descriptions), this study offers a new validity framework for vocabulary grading. It complements the guidance of the CEFR functional approach by outlining the lexical exponents needed to achieve the competences described in the framework, with the ultimate goal to increase the efficiency of language learners in achieving their communication goals.

17:15–17:45 **The construct of the collocation knowledge in language testing**

Margarita Alonso-Ramos

It is frequently stated that collocations are challenging to L2 learners, even to the advanced ones. This statement appears usually inside the literature which focuses on English. However, it is not obvious if the same situation is applicable to L2 learners of other languages such as Spanish. In order to verify it, it is necessary to design a collocation test which can be used to learners of Spanish. However, before implementing a collocational test, a reflection on the collocational construct is necessary.

I will review the different collocational constructs which are meant to be measured in some collocational tests administered to learners of English. As I will show, the concept of collocation which is used in these tests includes different phenomena ranging from binomial phrases to idioms, which means certain fuzziness of the construct of the collocation knowledge. I will defend that this construct is not necessarily fuzzy, but it has not been operationalised adequately. In answer to the question that Henriksen (2013) raises about the possible need to adopt another model for the knowledge and use of collocations, rather than the adopted for single words and formulaic sequences, the answer is yes: we need another model. This model must be inspired by the Explanatory and Combinatorial Lexicology (Mel'čuk 2012), because this is the framework that offers a more comprehensive vision of the phenomenon of collocations. Following this framework, I will finish by showing how it is possible to design a collocation knowledge test for learners of Spanish as L2.

19:00 VISIT TO THE CITY

SATURDAY 7th MAY

9:00-10:00 KEY NOTE PRESENTATION

Room: Conference room *(Ground floor NEXUS – Building 6G)*

Chair*: Claudia Harsch*

9:00-10:00 **Fluency and Interactivity**

April Ginther: *Associate Professor; Director of OEPP*

Fluency, and the variables associated with its representation, are of interest to a wide variety of researchers for at least two reasons: (1) the association of fluency with general language proficiency and (2) the relative ease with which temporal measures of fluency can be captured. Temporal measures of oral fluency (speech rate, mean length of run, number and duration of pauses, pause placement) have served as useful proxies for human raters' evaluations of spoken speech, and length variables (number of words, phrases, clauses) also serve as useful proxies for human raters' evaluations of writing. Reading rate variables (number of words/minute or syllables/second) also serve well as proxies for general reading proficiency. However, temporal measures of fluency are viewed as indirect and narrow – poor proxies of the richer representations of language proficiency that we value more highly. In this presentation, I will discuss the relationship between fluency and interactivity and will explore the ways in which the development of second language fluency is foundational to broader notions of communicative competence

10:00- 11:00 PAPER PRESENTATIONS

Room: Conference room *(Ground floor NEXUS – Building 6G)*

Chair*: Claudia Harsch*

10:00- 10:30 **Construct(s) measured in face-to-face and video-conferencing delivered speaking tests**

Fumiyo Nakatsuhara, Chihiro Inoue, Vivien Berry, Evelina Galacz

This presentation reports on a two-phase research project that explored the use of video-conferencing technology to deliver and conduct the face-to-face version of the IELTS Speaking Test. Phase 1 consisted of a small-scale investigation involving 32 test-takers and 4 examiners; Phase 2 was larger-scale with 99 test-takers and 10 examiners.

The two studies were designed to compare the construct(s) measured in the video-conferencing and face-to-face modes, by examining (a) test-takers' scores and linguistic output on the two modes and their perceptions of them, and (b) examiners' test management and rating behaviours across the two modes. Phase 2 of the research also investigated (c) the effectiveness of examiner training for the video-conferencing delivered test that was developed based on Phase 1 findings.

In both studies, test-takers took two IELTS Speaking Tests under face-to-face and video-conferencing conditions. A convergent parallel mixed-methods design was employed, triangulating multiple sources of information such as test-takers' scores, language functions elicited and feedback interviews. Examiners completed a feedback questionnaire and participated in retrospective verbal report sessions and focus group discussions to elaborate on their behaviour as interlocutors and raters. All test sessions were observed and field notes were taken.

The results suggest that the speaking construct remains essentially the same across the two modes, although some differences were observed in test-takers' functional output and examiners' interlocuting and rating behaviours. In the presentation, we will discuss the comparability of the

two modes and the implications, as well as caveats, relating to the use of the video-conferencing mode.

10:30- 11:00 **Test-taker interaction: How do raters perceive the construct?**

Linda Borger

The paired speaking test format is commonly used in both classroom-based and high-stakes assessment contexts, one of its advantages being the potential for eliciting a wider range of interactional functions than the oral proficiency interview. However, the construct of paired oral interaction is complex and thus poses challenges for rating. A major concern is how the co-constructed nature of the performance, and the 'unpredictability' that this brings about, affects the fairness of ratings. The objective of this presentation is to discuss data from a study exploring the construct of interaction in a paired speaking test from the rater's perspective. Thirty-one raters assigned holistic scores to six paired conversations from an authentic high-stakes EFL test and provided written verbal reports on features of the peer interaction that they reported contributed to their judgement. The written comments were segmented and coded. Findings indicate that the raters attended to individual features of each test taker's communicative ability, as well as to interactional features that were co-constructed by the two test-takers. Frequent comparisons were also made between the two candidates, indicating that the construct of spoken interaction was perceived as a joint achievement. Furthermore, raters reflected on the matching of candidates and how they performed in relation to one another. Examples from the different types of rater comments will be given, focusing on the issue of separability of scores for co-constructed interaction. Finally, implications for the development of rating scales, including the construct of peer interaction, will be considered.

11:00- 11:30 COFFEE BREAK

11:30- 13:30 PARALLEL PAPER PRESENTATIONS

Room 2.7 (2nd floor)
Chair*: Cristina Rodriguez*

11:30- 12:00 **Revisiting an EAP Speaking test: what is 'EAP Speaking', exactly?**

Bruce Howell

The TEEP (Test of English for Educational Purposes) is a long-standing EAP (English for Academic Purposes) test, currently used as an exit test on 'pre-sessional' EAP courses. A new Speaking component introduced in 2010 has a two-phase structure of monologue plus dialogue. Design of this test came from a desire to operationalise a series of values underpinning existing teaching: a strong teaching materials-to-test link, a topic-based approach, pairing for genuine interaction, a clear interlocutor script which minimises anxiety, inclusion of planning time, and so on. The test designers considered as fundamental the context of examinees having to explain and discuss suitably challenging but also motivating 'academic' topics, for authenticity. The overlying principle is that the monologue plus dialogue arrangement replicates the experiences of speaking as a student in English, such as making a presentation and participating in a seminar discussion.
Given a growing test-taking population and increasing demand for quality assurance measures in UK universities, a deeper look into the test's validation is underway, using frameworks such as Weir's (2005) and Bachman and Palmer's (2010). At the forefront is defining the 'EAP Speaking' construct. What is EAP Speaking, and does TEEP Speaking test all relevant sub-skills? Are any omissions justifiable for practical reasons?

The aim of this paper is to exemplify on-going validation of a small-scale but high-stakes speaking test in a well-established teaching setting, by first revisiting the construct definition. Sharing experiences will benefit others in a similar environments, especially in European universities where 'EAP Speaking' has increasing importance..

12:00- 12:30 The construct validity of a C2 exam for university students

Laura Riera Grau

Since its beginnings, the Language Service at the *Universitat Autònoma de Barcelona* (UAB) has been concerned with developing a quality exam system to certify levels of language proficiency in terms of those described by the Common European Framework of Reference (CEFR). Until fairly recently Exams ranged from A1 to C1, with the B1 and B2 levels being those most widely assessed in terms of the number of test-takers. However, in the last five years, a change in tendency has been observed. Undergraduates now arrive at university with a higher level of English than before, and as a result, demand for a C2 university exam has increased. In September 2015, in the light of this growing interest, the Testing Unit at the UAB Language Service was given a two-year brief to design and develop an exam aiming to assess the C2 level as described by the CEFR.

In this presentation, I will discuss the concerns of the team members in terms of content validity in order to produce a competence-based exam with relevant, representative tasks of the C2 descriptors, which allows us to make valid and reliable generalizations about the proficiency level of the test-takers in terms of the CEFR levels. I will outline the steps taken in this initial design stage as well as presenting some of the problems encountered in the process. Finally, I will discuss the steps planned to validate our construct.

12:30- 13:00 Writing on Admissions Tests and in University Classes: Two Constructs or One?

Brent Bridgeman.

Many tests used for university admissions contain a writing component. The assumption is that these short writing samples obtained under strict time limits are a reasonable proxy for the construct of interest—the ability to write coherently and effectively in university courses in which extended writing is done without time limits and with the opportunity to produce several drafts a final product. In this study, we asked graduate students who had taken the GRE revised General Test (GRE) to submit copies of two course-related writing samples. We asked that the submitted papers be approximately ten pages or fewer in length; they could be essays, term papers, or book reports, for example, but not very brief documents such as poems. The samples were scored on a 0-6 scale by university faculty members using a scoring guide that reflected a concept of critical thinking characterized as indicative of "scholarly habits of mind." Rater reliability was .70 and task reliability (based on the correlation of the two independent writing samples) was .54. The operational scores from the GRE Analytical Writing (AW) test are based on two thirty-minute writing samples. The correlation of the GRE AW scores with the university writing samples was .35. We divided both the AW scores and the university sample scores into high and low categories (high = 5.0 or above; low = 3.5 and below). Only 4% or the students who were low on AW were high on the university tasks while 29% were low on both.

13:00- 13:30 Assessing oral proficiency in a foreign language

Natalia Ringblom

This presentation discusses the complexity of measuring oral proficiency in an oral test situation at the university level. This difficulty seems to be due to a large number of variables that interact with each other, as well as the difficulty of assessing each and one of them when the holistic impression about the student's performance should be created.

The present study reports a small-scale project that was set to estimate the level of proficiency of the students at the oral language tests. 35 students of Russian as a foreign language participated in this study. A corpus of spoken tests (a total of 375 minutes) produced by the beginning learners of Russian were scored for both analytic and holistic features such as: (1) pronunciation, (2) lexical knowledge, (3) fluency, (4) grammatical complexity, (5) accuracy and (6) general impression. The purpose of the analysis was to examine the relationship between the grade given and the different variables.

A detailed linguistic analysis of the recorded tests was carried out and compared with the marks given by the teachers. Following these comparisons, the interviews with the teachers were made. The results indicate that in the absence of established and accepted by all the teachers guidelines of the students' proficiency, the teachers used their own subjective criteria when giving the grade (cf Chambers & Richards 1993). The results of this study call for use of caution when decisions about individuals are made based on purely subjective criteria.

Room 2.8 (2nd floor)
Chair*: Elaine Boyd*

11:30- 12:00 **The CEFR as construct across national contexts – Is your B2 my B2?**

Franz Holzknecht, Ari Huhta

Despite its widespread use as underlying construct for language assessments, studies comparing how the Common European Framework (CEFR) is operationalized across different national contexts are scarce. This study addressed this need by cross-validating CEFR-linked rating scales for writing across two European countries.
One hundred performances of teenage test takers based on two different writing tasks were collected in both a central European and a northern European country. A team of trained raters in both countries applied their own national CEFR-linked rating scale to all of the 200 performances. Each script was rated by at least three raters within each team. The ratings were analysed with FACETS.
Although there was overlap in the ratings at certain CEFR levels, inconsistencies between the two groups of raters were observed. The results show that, despite the fact that different countries use the *Common* European Framework as construct in their curricula, they seem to interpret the CEFR descriptors slightly differently, depending on issues such as scale point (lower vs. higher levels) and the type of scale (holistic vs. analytic). The results also shed light on the role of the task in defining the construct, as differences were observed relating to the task test takers performed.

Studies of this kind should thus help stakeholders across national contexts gain a better understanding of what is being assessed in order to make more justified decisions about test takers' competences. The study highlights the need for more investigations on how the CEFR is operationalized as construct across national contexts.

12:00- 12:30 **Exploring interactions between learner and task characteristics in a reading test of French for young learners – the example of the language of rubrics and response**

Katharina Karges, Malgorzata Barras, Peter Lenz

The study we are reporting on is intended to inform test item development for an upcoming computer-based large-scale assessment of 6th graders in Switzerland. The assessment concerns the students' first foreign language (French, German or English, depending on the region).

One of the item features in question is the language of the questions, options, and/or short answers, depending on the item type. Should, in the German-speaking region, this language be German, the language of schooling, or French, the target language of the reading test? Which language would lead to less construct-irrelevant variance? This question was explored from a qualitative as well as a quantitative angle. Qualitative information was gathered by means of retrospective interviews with individual students, as well as questionnaire items for over 500 students involved in the reading survey. For the reading survey, 36 French reading items were

prepared in four variants each: MCQ+German; MCQ+French; SAQ+German; SAQ+French. In addition, 18 matching items were presented in both language versions. Each student solved a balanced subset of these items. In addition, they took a series of tests focusing on specific cognitive and linguistic component skills and knowledge. The aforementioned questionnaire was used to gather more potentially relevant information, e.g. on motivation and attitudes. The data thus available allows for detailed analyses on the impact of specific tasks features, such as the language(s) used, on different types of test-takers.

12:30- 13:00 **What is speaking made of? Comparing 9th graders' speaking performances in English and Swedish**

Raili Hilden, Marita Härmälä

The ability of speaking comprises multiple competences ranging from general declarative to linguistic knowledge and skills. In the study at hand, we scrutinize and compare the competence structures of speaking as produced by pupils at age of 15.
The data derive from a national evaluation of learning outcomes carried out in 2013. A sample of 1 500 pupils of English and 859 pupils of Swedish performed four speaking tasks: a monologue, two dialogues and a paired discussion task. The tasks were broadly targeted at level A2. The performances were video recorded and assessed by teachers. Then, 10 % of them were censored by external raters. The correspondence between the two ratings and was considered sufficient (0.70-0.80). In further analyses, the speaking scores were correlated on one hand (1) with the skills of listening, reading and writing, and on the other, (2) with sets of study practices and attitudes. The correlations with reading, listening, and writing skills stand for linguistic competences and the connections with study practices and attitudes reflect the role of general competences. The interdependence of the components was corroborated by classical regression analysis.

The results indicate some differences in the construct structure of speaking across the two languages. The monologue structure is more similar in both languages, whereas the dialogue performances in English mirror more pupils' out-of-school practices and everyday contacts with English. In Swedish, the study practices at school seem to explain more the construct of speaking.

13:00- 13:30 **Construct validation in a test of reading for young learners**

Angela Hasselgreen, Torbjorn Torsheim

All schoolchildren in Norway are subject to national tests of English reading at the start of 5th and 8th grade. The construct of English reading is defined in the test, based on elements of the school curriculum, as various aspects of reading comprehension (involving finding specific information and main points in texts of varying length, and inferencing), as well as understanding vocabulary and grammar in context.
The test items are designed to cover all these features of the construct, with each item being mainly associated with a particular feature. Moreover, the way test results are to be interpreted by stakeholders assumes a progression in the difficulty of items targeting specific features, based on the levels of processing in Khalifa and Weir's (2009) model of reading.
The study presented in this paper aims to establish:
1) [T1] the extent to which item content features predict variation in item difficulty in battery of 200 English reading items.
2) the extent to which this prediction of difficulty supports the underlying construct and the theoretically-derived manner in which the tests scores are currently interpreted, using band scale descriptors.
As well as presenting the theoretical basis for and findings of the study, the paper will give an account of the processes followed. These include operationalisation of item features, item feature classification and scoring of approximately 200 items, rater-reliability assessment, and finally regression model specification and testing. The estimated regression weights analysis provides information about the extent to which item features predict item difficulty.
[T1]This is a BIG question, which we might not be able to address in this study,

although the result are of relevance.

12:00- 13:30 SYMPOSIUM

Room: Conference room *(Ground floor NEXUS – Building 6G)*

Revisiting the speaking construct: Multiple perspectives

Organizers: *India C. Plough* (Michigan State University)
Jayanti Banerjee (Worden Consulting LLC)

Presenters: *Spiros Papageorgiou* (Educational Testing Service)
Cathy Taylor (Trinity College London)
Alistair Van Moere (Pearson)

Discussant: *Steven Ross* (University of Maryland)

The notion of construct is central to the endeavor of developing and implementing valid assessment instruments and procedures. Construct definition entails specification of the test purpose, its target test taker, and the claims that will be made based on the test results. This symposium will address these key construct issues with respect to the assessment of speaking in three quite different traditions, representing the range of possibilities in test administration, content, and scoring.

The **speakers** will be:

Alistair Van Moere (Pearson), who will present a fully automated test of speaking that is part of the English for Professionals Exam (EPro™). EPro™ is designed to measure how well a person can understand and communicate in workplace English. During the speaking test, speech is elicited in seven item-types that evaluate test-taker proficiency under different cognitive and communicative demands.

Spiros Papageorgiou (Educational Testing Service), who will examine the speaking section of the TOEFL iBT Test, a test of university-level academic English that targets a construct of communicative competence defined in terms of the domain(s) of language use, the goals of communication, and the language abilities needed to achieve communication goals. This construct is operationalized through open-ended computer-mediated speaking tasks that require the integration of written and oral academic content and is scored asynchronously by human raters.

Cathy Taylor (Trinity College London), who will discuss the Graded Exams in Spoken English (GESE), a suite of 12 exams which are a face-to-face interview between an examiner and a candidate. The exam simulates real-life interactions in which the candidate and the examiner exchange information, share ideas and opinions. GESE provides a measure of linguistic competence from absolute beginner (below A1 CEFR) to full mastery (C2 CEFR).

Each presenter will speak for 18 minutes. After providing a very brief overview of the test they will explain the theory of language learning/acquisition that informs the test design, highlighting specific predictions and assumptions and how they are operationalized. Finally, they will discuss learner (test) performance in terms of the implications for theory and pedagogy.

Steven Ross (University of Maryland), the **discussant**, will speak for 20 minutes. With his expertise in Second Language Acquisition and assessment he will critically examine the theoretical underpinnings of each test, identify commonalities between tests, and discuss the implications of the differences.

The symposium will then be opened to the audience for their comments and questions (16 minutes).

13:30- 15:00 LUNCH BREAK

15:00-16:30 ANNUAL GENERAL MEETING
Room: Conference room *(Ground floor NEXUS – Building 6G)*

16:30- 17:00 COFFEE BREAK

16:30- 17:00 POSTER PRESENTATIONS (Ground floor-NEXUS- Building 6G)

17:00- 18:00 PARALLEL PAPER PRESENTATIONS

Room 2.7 (2nd floor)
Chair*: Slobodanka Dimova*

17:00-17:30 **WORK IN PROGRESS**
Including Teacher Voices Through An Institutional Writing Criteria
Dilek Salki, Bahar Hasirci

The aim of this presentation is to explain the reasons for and the process of improving an analytical writing criteria in a preparatory programme of an English-medium university in Turkey. The institutional criteria had been previously developed using a theory-based approach, including the writing features (content, organization, grammar, and lexis) which form the paragraph and essay writing construct in our context. However, in time, feedback collated from instructors and observations during writing standardization sessions revealed that there were problems in the following areas: the descriptors did not clearly define the expected features of the writing skill and did not help to differentiate between bands easily; certain elements of the writing construct were missing; and the descriptors in the criteria did not necessarily reflect instructors' justifications for assigning a certain grade. All of these revealed that there were mismatches between instructors' understanding and expectations of the writing skill compared to the descriptors in the criteria. The improvement process took on a data-driven approach by asking instructors to grade student samples and justify their grades. The descriptors in the revised criteria were grounded on the commonly expressed justifications of instructors. By the end of this process, we were able to create an improved version of the writing criteria. However, we also realised that instructors had different ways of interpreting the same criteria and there was need to construct a more common institutional understanding through a new approach towards standardization and training sessions.

17:30–18:00 **An examination of how analytic raters adapt to holistic marking: a quantitative study**
Judith Fairbairn

The focus of this research is on how raters with a background in analytic rating adjust to marking holistically. What is of interest to the researcher is if raters internalise a particular rating construct from a test, which inhibits their ability to mark using a different marking system with a different construct.

Research shows that good training and marking scales can mitigate rater effects. It therefore seems possible to cultivate a rating culture in the cognitive style appropriate for the construct of a test. What has not been studied is if raters belonging to a rating culture can easily switch to another rating culture. Does a rating culture form around a test construct and its marking scales that is difficult to change? Do raters need more time to train if they are accustomed to a different rating culture? Do they require more frequent standardisation and monitoring if they are using more than one marking construct in their rating work?

In this paper, the analytic rating culture background variable is studied to determine if raters can reliably switch to holistic marking. Outcomes from this research can help improve rater

Room 2.8 (2nd floor)
Chair*: Marisa Carrió*

17:00-17:30 **Classroom Assessment Construct: EFL Teachers' Perceptions and Practice**

Hossein Farhady

Classroom assessment (CA) is assumed to be grounded on certain principles such as the centrality of learning, learner engagement in assessment process, and the significance of feedback to improve teaching and learning process (Purpura & Turner, 2024; Purpura, 2015). Such principles have placed more demands on classroom teachers to effectively implement CA (Inbar-Laurie & Levi, 2015; Xu, 2015). Most of the desirable characteristics of CA have not have been translated into the working knowledge of EFL teachers who are the actual agents of implementing these principles. Therefore, it would be revealing to collect information on teachers' perceptions of CA and the underlying structure of CA construct from their perspective. To address the issue, a questionnaire was administered to 120 EFL teachers in Turkey. The major dimensions of the questionnaire were a) to check EFL teachers' perceptions of CA, b) to identify the methods, tools, and strategies they use in practicing CA c) to identify what additional tasks they think they should practice but they do not or cannot, and d) to find the influence of teachers' perceptions of CA on their actual practices. The preliminary findings reveal that EFL teachers have positive attitude towards CA and they attempt to implement some of its principles. However, they claimed that they were aware of the insufficient knowledge of these principles and they showed great interest in receiving professional help in this regard. The results of the study will be discussed in detail and the implications for teacher education programs will be explained.

17:30–18:00 **Test-takers' voices in assessment tasks**

Sehnaz Sahinkarakas

Recent trend towards social-psychological issues in language learning, specifically individual differences, inspired some scholars in the field of language testing and assessment. McNamara is one of them who argues that we need to rethink validity theory to correspond to the needs of learners and individual differences, which has not been well considered by language assessment theories. In line with this argument, this case study is an attempt to discuss one way of including students' voices in assessment tasks. The participants of the study were 75 ELT Freshman students taking contextual grammar course. Before the mid-term and final exams, the students were asked to write the assessment tasks (content and method) they wanted to be assessed and to explain why they wanted these tasks. These student-specified test objectives were used for the development of some of the assessment tasks. Other tasks which were not mentioned by any of the students but covered in the course syllabus were also developed. The scores in the student-specified assessment tasks and the teacher-decided ones were subjected to t-test and Pearson correlation coefficient to analyse the relationship between them. At the end of the exams, nine students were interviewed to reveal to some extent the validity of student-stated objectives and their opinions about involving them directly in the testing process. The results may provide valuable insights into the way we conceptualise the constructs and how these constructs correlate with test-takers' needs.

Room 2.9 (2nd floor)
Chair*: Norman Verhelst*

17:00-17:30 **Tests of lexicogrammar: using expert judgements to investigate their underlying construct**

Theresa Weiler

This presentation reports on a study that investigated the construct underlying lexicogrammar tests used in a high-stakes, school-leaving test context. More specifically, it explored what is being tested by items organized in four different item types: multiple-choice gap-fill, banked gap-fill, word formation gap-fill, and editing tasks. To this end, a mixed-methods approach was employed, combining expert judgements, test-taker introspection and test performance data. This talk will primarily focus on the findings from the expert judgements, and also the use of this particular method to gain insights into the test construct and inform test validity.

A group of 16 highly-experienced language teachers and testers was asked to evaluate what each item in three versions of the lexicogrammar test was testing. To assist the experts with their task, a judgement grid was designed, based on Purpura's (2004) model of grammar. For each item, the judges had to indicate in the grid what they thought it was testing. Afterwards, they were asked to comment in a questionnaire on their judgement task experience. The resulting data were analysed in terms of (1) inter-judge reliability, (2) judges' views on the construct covered by the items, and (3) feedback on the judgement task. The talk will report and discuss a) how the judgements helped describe the construct operationalized by each of the four item types, b) to what extent the item types lend themselves to testing the same/different aspects of lexicogrammar, and c) the use of the judgement research method for construct investigation and test validation.

17:30–18:00 **Psychometric Analysis: Evaluation of EAP Language Tests items and a Pre-sessional Course**

Ricardo de la Garza Cano

Designing language test items is a process in which test developers need to take different aspects into consideration and one of these is the test-takers' personal characteristics such as gender, academic background, first language and nationality (Kunnan, 1995, 1998 and 2000). This study reports on the results conducted through item analysis to scrutinise the effectiveness and quality of items in reading and listening assessment components and linguistic features from speaking and writing components but also to identify any relationships between the test takers. This research aimed to i) identify the quality and effectiveness of the items, ii) measure the efficiency of the distractors from multiple choice items, iii) find any relationships between the items with the test-takers' performances, and iv) briefly analyse the Pre-sessional course through a focus group of test-takers' perceptions about the structure, methods of teaching and learning, assessment and testing environment. The test takers were analysed through their personal characteristics proposed by Bachman and Palmer (1996). The first analysis is between test-takers' performances and genders. The second was based on their academic fields. Three academic disciplines were created based on the test-takers' degrees and the last one was based on their nationality. The methods of this study were difficulty index, discrimination index and distractor efficiency. In addition, the overall means from the three sections of test-takers' background was calculated to find relationships to the items and the performances. The study provides clear implications for EAP test developers as well as for the development of EAP courses.

20:30 **CONFERENCE DINNER & MUSIC – DANCE**
(Hotel ASTORIA PALACE – Plaza Rodrigo Botet 5– *City Centre*)

SUNDAY 8th MAY

9:30- 11:00 PAPER PRESENTATIONS

Room: Conference room *(Ground floor NEXUS – Building 6G)*

Chair*: John De Jong*

9:30- 10:00 Re-defining the construct of vocabulary size tests: Challenging Conventions

Benjamin Kremmel.

What are vocabulary size tests assessing and what do their scores mean? Although it seems straightforward, the construct of vocabulary tests is based on a number of assumptions that do not seem empirically motivated, but nevertheless have become unquestioned conventions. This talk will problematize and challenge two key assumptions, and suggest a re-definition of these aspects of the construct to make vocabulary tests more diagnostically useful.

The first convention addressed is the counting unit. Most vocabulary size tests have used word families, but this may not be the best unit. Even if learners know one or more members of a word family, they do not necessarily know all of its members. The study presented will explore this by comparing 99 EFL learners' knowledge of root forms with their knowledge of the entire word family. The findings suggest that word families are an inappropriate counting unit as learner only managed to make the connection between root form and derivatives in 73% of the cases. The talk will therefore give reasons why the lemma would be a more suitable counting unit.

The second assumption is that the vocabulary continuum should be divided into bands of exactly 1,000. However, current research indicates that a one-size-fits-all approach to frequency division may not be the best solution. Discussing coverage research on four different corpora, the talk will argue that more narrowly defined bands (500 units per band) might be more useful for higher frequencies, while wider bands (2,000) might be perfectly workable for lower frequencies.

10:00- 10:30 Investigating the construct of speaking proficiency for young language learners: A discourse-analytic study.

Ching-Ni Hsieh, Yuan Wang.

Consistent with this year's conference theme "*Assessment of what?... – Revisiting the issue of construct(s)"*, this study examined how linguistic features characterized the speaking proficiency of young language learners. We investigated the link between observed test performance and the underlying constructs of speaking ability, within the context of the speaking section of a large-scale young learner language assessment. The construct of speaking proficiency for young learners has not been widely explored (McMay, 2006), and further studies are critically needed for meaningful score interpretations of young learner assessments. In this discourse-analytic study, we used a total of 358 speaking samples from 179 test takers who responded to the picture-narration task and the integrated Listen/Speak nonacademic task of the speaking test. Test takers, who were classified into four speaking proficiency levels based on their test scores, were from a variety of first language backgrounds (mean age = 13.60 years). Three conceptual construct categories assessed by the test were analyzed,*delivery*, *language use*, and *content*, and 22 linguistic features tapping into the three categories were identified for analysis. The linguistic features were analyzed using computer software and human coders. Results of a series of repeated-measure ANOVAs indicated that the majority of the 22 linguistic features differentiated test takers across proficiency levels, with the features of fluency, vocabulary, and content quality having the strongest impact. We

concluded that the findings of the study would extend our knowledge of the speaking constructs for young language learners, and have implications for young language learner assessment.

10:30- 11:00 **What to test at C1?**

Susan Sheehan

This paper addresses the conference theme of constructs in assessment by seeking to identify criterial features at C1. It reports on a project which aimed to identify key criteria of written and spoken English at C1. The project hoped to bring clarity to our understanding of an under-specified and under-described level (Weir, 2005, Green, 2012). This new understanding could aid test developers when creating tests. The project was an extension of the British Council / EAQUALS Core Inventory for General English (Core Inventory). It aimed to provide data-based evidence conducted with learners to support existing theoretical work.
Methodology
MA TESOL students with IELTS scores of 6.5 or above were invited to take a test of written and spoken English. The test had been created to satisfy the requirements of an external validation agency. The scripts were analysed with ATLAS ti software to identify which of the features described as core in the Core Inventory are found in the scripts and with what level of frequency. The software was used on both the written and audio data
Results and Conclusions
The language points which could be considered criterial tended to be those relating to argumentation and expressing feelings and attitudes precisely. There is some evidence to suggest that giving advice could be considered to be criterial. Perhaps the significance of the project lies in the creation of an approach to identifying criterial features by using the Core Inventory.

11:00- 11:30 **COFFEE BREAK**

11:30-12:30 **ROUND TABLE & CONCLUDING REMARKS**
Room: Conference room *(Ground floor NEXUS – Building 6G)*
Topic: *Summing up, thinking forward*
Moderator: *Claudia Harsch*
Speakers: *Neus Figueras, April Ginther, Constant Leung, Sauli Takala*

12:45 **CLOSING CEREMONY**
Room: Conference room *(Ground floor NEXUS – Building 6G)*

13:00 **TRIP TO *LA ALBUFERA LAKE***

TRIP to *La Albufera* Lake, for all participants by request.
Picnic lunch.

Arrival to the lake and short explanation about the area and traditional houses. Tasting of typical Valencian food during the afternoon snack and tour in a small fishing boat.

Estimated arrival to Valencia at 17:00

POSTERS

POSTERS will be exhibited during Coffee and Lunch breaks

SATURDAY 7th	POSTER PRESENTATIONS
16:30-17:00	**A framework for enhancing teachers' assessment literacy** *Dina Tsagari, Karin Vogt, Ildikó Csépes, Tony Green, Nicos Sifakis* This poster will present the aims of a three-year long project involving a diverse network of experts from different European countries who aim to develop an effiecient and sustainable LTA training infrastructure primarily for English language teachers to help them develop sufficient assessment literacy skills. The project entitled 'Teachers' Asssessment Literacy Enhancement (TALE)' aims to contribute towards: - the development of innovative training materials and services that will primarily be delivered through online training systems in synchronous and asynchronous modes with taught and self-access options; the proposed training system is expected to offer continuous support and mentoring to teachers in the countries involved in the project with the intention to roll out the services across Europe by the end of the three year period; - an innovative approach to the sharing of the LTA expertise between European educational contexts that takes advantage of web-based collaboration tools; -collaboration between and within disciplines and between various training sectors in order to foster efficient and meaningful assessment suitable for language learners in primary and secondary education. **Analysis of Reading Construct in PTE Academic** *Abdullah Arslan, Salih Ozenici* Closely associated with skills that students need to gain in order to deal with the requirements of high-stakes tests in English (Alderson, 2000; Bachman and Palmer, 1996), construct validity focuses on the structure of a test as well as the reflection of the psychological reality of behaviour in the area being tested through a test (Hamp-Lyons, 1990). Launched in 2009, the Pearson Test of English Academic, which is one of the public tests recognized by institutions and programs across 56 countries worldwide, is said to assess the English language skills of various cohorts for the required academic study, professional recognition or all visa classes in Australia (UK Higher Education Institutions Information Pack, 2015), for this reason this study explores the construct of reading test underlying PTE Academic through the taxonomy proposed by Weir and Urquhart (1998). In this context, the corpus of reading tasks including 4 complete tests taken from Practice Tests Plus with keys (Pearson Education, 2013) is employed to conduct an analysis of the reading tasks based on the two dimensions of the analytical framework–level of engagement and type of engagement in the taxonomy of Weir and Urquhart (1998). A general and detailed accounts of the analyses and their results will be presented followed by the identification of each of the task types of reading test module in the PTE Academic, along with the discussion how each relates to the 'level of engagement – type of engagement' dimensions. **Motivation and attainment in English for tourism among university undergraduates** *Zoltán Lukácsi* Obtaining a state-accredited language exam at level B2 is an academic requirement in Hungary for university undergraduates working towards a Master's degree. Since mock

exams are rarely available, live administrations incur considerable costs and inflict stress; the Budapest Business School developed a complex departmental test of English for tourism in 2013. Students take the Listening, Reading and Use of English papers on the outset and in the end of language instruction and are informed about their progress and attained level with regard to a set standard. Exam data are processed in OPLM (Verhelst, Glas & Verstralen, 1995) and true score equating is applied for consistency of requirements.
Given the role of motivation in foreign language learning and the fact that a third of the students fail to achieve the desired learning outcomes, the university felt the growing need to collect information about this background variable. In October 2015, 244 freshmen completed a previously validated motivational questionnaire (Török & Csizér, 2007) based on Dörnyei's (2005) views on motivated language learning behaviour and self-perceptions. The data from the 63 Likert scales were analysed for factor structure and relationship with test paper scores. Principal axis factoring with Varimax rotation confirmed the original 11-factor solution with a markedly weakened written communication component. The results also showed that Listening and Use of English were significantly correlated with (a) oral communication in English, (b) L2 self-assessment, (c) motivated learning and (d) attitudes towards US citizens, whereas Reading was only associated with integrative motivation.

Choice Given or Not?: Assessing Writing Skills of EFL Students
Meral Melek Unver.

Writing, as being a productive skill, is regarded a challenging skills for language learners (Graham, Harris&Mason, 2005). It is mainly because writing requires learners to have control over the mechanics of the language, to have sufficient grammatical and lexical knowledge, to be creative and to have topical knowledge of what is being asked (Wall, 1981). Writing is especially difficult for English language learners in a non-English speaking country, like Turkey because they have to explore thoughts and ideas and put them in words in an accurate, coherent and communicative way. This study aims to explore how English language learners react to writing tasks on a written performance examination in which students are provided two alternative items with at least three prompts and asked to write about only one of them, which is thought to provide students with relatively less challenging tasks considering the time pressure and the complex nature of writing in a foreign language. However, do the test takers really consider it an exam with choices provided? What makes them to pick one of the alternatives and write about them? Do the prompts given really help them produce ideas easily under time pressure? The ultimate aim of this study is to seek answers to the above-mentioned questions in a preparatory program of a Turkish state university through the examination of 800 exam papers and semi-structured interviews with randomly-selected students. The results will give insights into the perceptions of the students of written exams and help design better tests.

The interface of consequential validity and language assessment literacy
Dina Tsagari, Karin Vogt

Theoretical approaches to validity (Messick 1989) and, in this connection, the argument-based framework for validity (Kane, 2006, 2013) have been in the centre of scholarly attention for a while.
More recently, Bachman and Palmer (2010) contend that the intended consequences of a test should be the starting point for test design, highlighting the need to justify test use rather than test score meaning (Chapelle, 2012). Kane (2013: 46) argues for including evaluations of specific kinds of consequences under the term validity. The consideration of the social consequences of test use, for impact, bias and washback (Davies, 2012) has been labelled consequential validity (Messick 1989).
When applying consequential validity to classroom-based language assessment, teachers seem to be important stakeholders (Chalhoub Deville, 2015). However, studies on washback (e.g. Cheng, 2008; Tsagari, 2009; Wall, 2005) have shown that consequential validity is at

risk when teachers' language assessment literacy is not developed enough, e.g. when teaching to the test takes place.
The purpose of the poster is to show the need for solid levels of language assessment literacy (LAL) for all relevant stakeholders (following e.g. Kremmel & Harding 2015), not only teachers, based on a sustainable construct of LAL. Chances and limitations of matching consequential validity and LAL of several stakeholders will be discussed as well.

Investigating the construct of a writing test across languages: a corpus based approach

Michael Maurer, David Moreno

Functional competence is "concerned with the use of spoken discourse and written texts in communication for particular functional purposes" (Council of Europe, p. 125). For the assessment of writing, "macrofunctions" are particularly important. These are defined as "categories for the functional use of […] written text consisting of a (sometimes extended) sequence of sentences" (Council of Europe, p. 126). Those macrofunctions are often included in test specifications in order to define its test's construct. However, the assessment of learners' functional competence as described in the Common European Framework of Reference (p. 125 ff.) poses certain challenges to language testers. The framework does not include an extended enumeration of macrofunctions, but only an unfinished list (Council of Europe, p. 126). In addition, currently there is a lack of specificity when it comes to translating functional competence into levels on the illustrative scales. More information on these two issues would be beneficial to language test developers in writing test specifications and defining a test's construct.
This poster presents the second stage of a project investigating data gathered from test takers writing samples. The learner corpora include the languages English, French, Italian and Spanish with a corpora size of over 400 texts per language. Stage one (presented at EALTA Warwick 2013) examined two languages (English and Italian) and two functions (explain and suggest). Stage two, presented in this poster, extended the languages under investigation to include French and Spanish and the functions analysed to include describe and narrate. All samples are based on standardized writing tasks developed for a national high-stakes exam. As part of the writing exam's construct, the tasks specifically target the macrofunctions listed in the CEFR. These aspects of the construct are analyzed with analytical software tools to answer the following research questions: Which macrofunctions are test takers capable of performing in different languages? Does the inclusion of macrofunctions in writing prompts mean that test takers actually perform these functions?

Construct validation: challenges of a multi-level Language Program

Yevgeniya Pronoza

The poster presentation will focus on practical ideas on how to deal with the challenges of operationalising the test construct in a multi-level English Language Foundation Programme. The programme serves the needs of over 5000 homogeneous language learners. The Assessment Unit I am heading is responsible for placing 3000 new comers every year into appropriate levels of the language proficiency, as well as assessing their learning at the end of various language courses. While test writing, the challenges we test writers face are as follows: construct underrepresentation, construct overrepresentation, individual interpretation of the construct by test writers, and test validation. There are other challenges as well but the focus will be on the most important ones. The presenter will share some practical ideas of what the Assessment Unit does in order to overcome the abovementioned challenges and adequately represent the construct outlined in the Test Specifications. Namely, the presenter will use both the achievement and placement tests used on the language programme as an example. Both types of tests are designed internally and are geared towards the needs of the English Language Foundation Programme. The

presenter will not be able to use the actual tests for security reasons but the sample test copies will be available. As for the Test Specifications, I will use the sections which are not confidential and are in open access on the programme website.