EALTA 2019

Transitions in Language Assessment

31 May – 2 June

Dublin, Ireland

EALTA www.ealta.eu.org European Association for Language Testing and Assessment

UCD DUBLIN

16th EALTA Conference

Organiser and Publisher

Applied Language Centre
University College Dublin
Belfield
Dublin 4
Ireland

Ealta2019@ucd.ie
https://www.ucd.ie/alc/ealta2019/

Photo credits

Page 17 © Main Restaurant, UCD
Page 19 © Fallon & Byrne, 11-17 Exchequer Street, Dublin
Page 62 © George Moore Auditorium, UCD
Page 63 © UCD campus map

Contents

Organising Committee

Chair

Anna Nunan

Scientific committee

Claudia Harsch
Peter Lenz
Jamie Dunlea
Sonja Zimmermann
Slobodanka Dimova
Zoltán Lukacsi

Conference committee

Joanne Dalton
Stergiani Kostopoulou
Christopher Pingeon
Brian Rice
Alex Runchman
Aoife Sheils

Vetters

Jayanti Banerjee
Tineke Brunfaut
Oscar Canela
John de Jong
Bart Deygers
Slobodanka Dimova
Jamie Dunlea
Gudrun Erikson
Neus Figueras
Doris Frötscher
April Ginther
Luke Harding
Marita Härmälä
Claudia Harsch
Peter Holt
Ari Huhta
Stergiani Kostopoulou
Benjamin Kremmel
Peter Lenz
Zoltán Lukacsi
Eli Moe
Fumiyo Nakatsuhara
Anna Nunan
Alex Runchman
Richard Spiby
Dina Tsagari
Aylin Unaldi
Norman Verhelst
Karin Vogt
Dianne Wall
Carolyn Westbrook
Sonja Zimmermann

President’s Message – EALTA 2019 Conference

Céad Míle Fáilte
A Hundred Thousand
Welcomes!

Dear EALTA members,

Céad míle fáilte – a hundred thousand welcomes to Ireland and to the 2019 EALTA conference on transitions in language assessment. Language assessment is a dynamic field that has undergone a number of changes in recent years. These include the emergence of different approaches to the assessment of language proficiency, changes in the design and delivery of assessment, as well as an increased need for assessment at different points of transition, be it for migration, employment or academic purposes. Such changes require flexibility on the part of test takers, teachers, educators and materials developers, but they also create a productive environment for research and practice. We look forward to exploring aspects such as new potentials of digitalisation, emergent questions of validity and fairness and the influence of a variety of societal and political factors on assessment.

Our three keynote speakers will address transitions in language assessment from different angles. Talia Isaacs, University College London, will present an overview of terms, themes and trends as well as a look at shifts in attitudes on the use of technology in assessment. Detmar Meurers, University of Tübingen, will explore issues related to computational linguistic analysis and language assessment. Michael T. Kane, ETS, will draw our attention to the validation of claims inherent in the interpretation of assessment results. The EALTA symposium this year addresses the issue of validity models in transition.

As in previous years, we are proud to offer a range of pre-conference workshops and would like to thank our workshop leaders for their input and commitment. This year there has been a strong interest in our six SIG groups on classroom-based assessment, CEFR, assessing speaking, assessment for academic purposes and assessing writing and migration and integration. For the second year, we can also offer a series of lunchtime workshops – many thanks to the colleagues offering these sessions.

On behalf of the Executive Committee, I would like to thank all those who have made this conference possible and particularly our generous sponsors, without whom we would not be able to offer such an excellent event.

Over the past year, based on your voices and feedback, we saw a number of important activities: webinars, SIG meetings and workshops supported by EALTA. Thank you all for your participation on behalf of EALTA. At the AGM, we will discuss further initiatives and look forward to your feedback on these. An election will be held this year for the role of Chair of the Committee for Conference Organisation. Please attend the AGM, your voice is very important to the future of EALTA!

The past year has also been my last year as EALTA president, and I would like to extend a heart-felt thank you to the Executive Committee and to all of you for your support and the trust you have put in me. In the past three years, we have implemented a number of initiatives, listened to your voices and needs, and addressed new challenges. I am handing over to Peter Lenz now, in full trust that he and the Executive Committee will continue to serve EALTA and the language assessment community.

Looking forward to a thought-provoking conference and a stimulating year ahead.

Claudia Harsch and the EALTA Executive Committee

Sponsors

BRITISH COUNCIL
ASSESSMENT RESEARCH GROUP

TRINITY
COLLEGE LONDON

OXFORD
UNIVERSITY PRESS

UCD
DUBLIN

IELTS™

Lancaster University

TGLS®
the global language system

ETS® TOEFL®

Panhellenic federation of language school owners
PALSO

centre d'idiomes
VNIVERSITAT DE VALÈNCIA

TestDaF
Test Deutsch als Fremdsprache

LRN
Learning Resource Network

Pre-Conference Programme Overview

Pre-Conference Workshops *Applied Language Centre*

Tuesday, May 28 (2pm) – Thursday, May 30 (noon)

Room 15	Room 12	Room 2
Many Facet Rasch Measurement: a practical guide using FACETS	IRT Analyses Using the Statistical Software R	Assessing speaking: Developing and Applying Rating Scales
Nahal Khabbazbashi & Tony Green	*Steffen Brandt*	*Kathrin Eberharter, Nivja de Jong, Jayanti Banerjee & Carol Spöttl*

Thursday, May 30

12.00-14.00 ***Early Registration*** *Applied Language Centre*

14.00-17.00 ***Parallel SIG Meetings***

Room 15	Room 3	Room 2	Room 12	Room 8
Classroom-based Assessment	***CEFR***	***Assessing Speaking***	***Assessment for Academic Purposes & Writing***	***Migration and Integration***
Dina Tsagari	*Neus Figueras*	*Kathrin Eberharter, Carol Spöttl & Nivja de Jong*	*Peter Holt, Claudia Harsch & Sonja Zimmermann*	*Bart Deygers*

Thursday, May 30

19.00-21.00 ***Welcome Reception***

Applied Language Centre (Daedalus Building)

IELTS
Official IELTS Practice Materials

Pre-Conference Workshops

(1) Many Facet Rasch Measurement: A practical guide using FACETS

Participants will be introduced to some of the issues associated with involving human judges in the rating process in speaking assessment contexts. They will become familiar with the basic concepts in Many-Facet Rasch Measurement (MFRM) and will learn how to run the FACETS programme and to enter, analyse, and interpret test data with more than two facets e.g. candidates, raters, scoring criteria, tasks, etc. Participants will learn to evaluate the extent to which raters are consistent in their rating behaviour and whether they display systematic differences that can affect candidate scores.

Content and method:

- Session 1 (afternoon of Tuesday 28th of May 2019) will cover some of the main issues related to rater-mediated contexts such as rater inconsistencies, relative severity levels of raters, systematic bias, and different tendencies in use of assessment scales. We will discuss some of the basic theoretical concepts behind MFRM, the practical contexts in which this type of analysis can be useful, and some rules of thumb for designing studies that use MFRM.
- Session 2 (morning of Wednesday 29th of May 2019) will cover the different steps involved in running MFRM with the programme FACETS using an example speaking test data set with three facets (examinees, rates, and criteria).
- Session 3 (afternoon of Wednesday 29th of May 2019) will focus on understanding and interpreting the various outputs from FACETS.
- Session 4 (morning of Thursday 30th of May 2019) consolidates previous sessions by providing participants with the opportunity to independently run, analyse, and interpret another data set. The session will end with a Q&A.

The sessions will be interactive with the first session starting off with paired/group work with a series of simple data sets and hand-on activities that will allow participants to get a concrete understanding of some of the issues related to rater-mediated contexts using an inductive approach. We will illustrate and demo the steps involved in running FACETS including preparing and formatting the data, creating control/specification files, defining models, and running FACETS. Following the demo, participants will be asked to try out all steps independently. At the end of session 3, we will ask participants to rate a sample speaking test using the CEFR scale. We will collate these scores which would then form part of the data used for the final session; an exercise that would encourage full involvement of participants from the original rating process to data preparation, data analysis, and interpretation. Participants will be asked to complete a series of tasks related to the data set individually and then to compare responses in pairs and small groups.

Workshop leaders:

Nahal Khabbazbashi

Nahal holds a DPhil in Education from the University of Oxford. She is a Senior Lecturer in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Before joining CRELLA, Nahal held the position of Senior Research Manager at Cambridge Assessment English where she led the research strands of a number of high-profile international projects in different educational and language testing contexts.

Anthony Green

Tony Green is Professor of Language Assessment and the Director of the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire. He is the author of *Exploring Language Assessment and Testing* (Routledge), *Language Functions Revisited* and *IELTS Washback in Context* (both Cambridge University Press). He served as President of the International Language Testing Association (ILTA) from 2015-16 and is an Expert Member of the European Association for Language Testing and Assessment (EALTA).

(2) IRT Analyses Using the Statistical Software R

This workshop will provide an introduction into the use of R, the R programming environment RStudio, and the TAM package, a powerful function package within R to conduct item response theory (IRT) analyses. Using these tools, we will go through all steps necessary to do an IRT analysis, and you will directly apply them either on language test data provided to all participants or on your own data set that you bring to the workshop.

Content and method:

- *Introduction to RStudio*
- How to load, save and run R syntax
- How to install additional function packages
- Tips & tricks
- *Introduction to R*
- Import and export of data sets
- Applying functions
- Re-coding data
- Data structures in R
- *IRT Analysis*

- Estimation: How to calibrate uni- and multidimensional Rasch models using TAM
- Model fit: How to decide on the dimensionality of a given test
- Item analysis, Part I: How to investigate the dimensional fit of items
- Item analysis, Part II: How to investigate the fairness of items across different groups using differential item functioning
- Person analysis, Part I: Differences between the possible estimates (WLE, EAP, …) and how to select the appropriate estimate for your analysis
- Person analysis, Part II: The advantages of models using background data (also called background or regression models) and how to implement them

The workshop will be fully hands-on. All participants will work on their own R analysis project and will conduct the presented steps directly on their own project using their own laptop computers (all participants need to bring a laptop). Individual problems or differences in the analyses will be shared and discussed across all workshop participants, so a large variety of possibly arising cases and resulting recommendations will be shown, and all participants will benefit from the discussion of these cases and get experience beyond the cases currently occurring in their data set but possibly occurring in future situations. To do so, the workshop is accompanied by a Google Hangout Session, allowing participants at any time to share their screen and show their current results. Also, the Google Hangout will be used to share code snippets and examples at any point in time.

Workshop leader:

Steffen Brandt

Steffen currently works for the Goethe University Frankfurt teaching at the Center of Methods in Social Sciences and is a consultant in empirical educational research. He works on the support and implementation of projects on behalf of universities and public institutions and has worked on the implementation of workshops and advanced training on the topics of item response theory, data science and R at the universities of Bamberg, Mannheim, Frankfurt, Magdeburg and Kiel.

(3) Assessing Speaking: Developing and Applying Rating Scales

This workshop will address the theoretical and practical underpinnings of rating scale development and design as well as the challenges of training raters to reliably apply the scales in testing contexts. While we will be focussing on speaking in this workshop, there will be opportunities to discuss how the principles of scale development and rater training apply to the context of writing. The target audience is graduate students, pre- and in-service language teachers and teacher trainers.

Content and method

- Session 1 Plenary: The Construct of Speaking
- Session 2 Plenary: Principles of Rating Scale Design
- Session 3 Workshop: Rating Scale Design
- Session 4 Discussion/Q&A: Rating Scales
- Session 5 Plenary: Principles of Rater Training
- Session 6 Workshop: Rater Training (Part 1)
- Session 7 Workshop: Rater Training (Part 2)
- Session 8 Discussion: Summary and Reflections

The workshop will combine the plenary delivery of input with small group discussions and hands-on activities in which participants will be involved in practical work, such as discussions of rating scales and descriptors, rating performances, and discussions of ratings. Participants are requested to bring a laptop computer so that they can work with data and save the results of their work.

Workshop leaders:

Kathrin Eberharter

Kathrin is an active member of the Language Testing Research Group (LTRGI) at the University of Innsbruck and is involved in a number of research projects. She is working on her PhD in the area of assessing speaking and a study investigating language acquisition in healthy subjects and patients with Multiple Sclerosis. She has taught the pre-service teacher training module on Testing and Assessment at the School of Education (University of Innsbruck) and is very experienced in rater training on teacher education programmes.

Nivja de Jong

Nivja is an Associate Professor in Second Language Acquisition and Pedagogy and chairs the Language Learning Resource Centre at Leiden University. Her research focuses on assessment and pedagogy of L2 speaking. From 2008-2012, she was principal investigator in a research project (funded by Pearson language testing and Utrecht University) on speaking fluency. In 2016, she taught on assessment of the productive skills in the EALTA Summer school. Since 2017, Nivja de Jong is an expert member of EALTA.

Jayanti Banerjee

Jayanti Banerjee has more than 15 years' experience in language testing and assessment. She teaches language testing and assessment courses at the Master's level and has supervised numerous Master's and PhD level research projects. She also has extensive practical language testing experience. She has managed a medium-sized language assessment programme. In this role she was responsible for the day-to-day running of the programme, including rater training and monitoring. Jayanti's recently published research includes a special issue of *Language Testing* focused on the assessment of interactional competence through speaking tests.

Carol Spöttl

Carol Spöttl is part of the Language Testing Research Group (LTRGI) at the University of Innsbruck, which is involved in language assessment research projects with partner institutions in the UK, Finland, Germany, Italy and Spain. Previously Carol led a government-funded project to reform the Austrian school-leaving exam for the foreign languages English, French, Italian and Spanish in upper secondary schools (2007-2013). She led the team that developed national assessment scales (A2-B2) for speaking and also produced the illustrative benchmark performances. Together with Kathrin, she led national workshops to train teachers in the use of the scale, interlocutor behaviour and speaking task design. Carol is a member of the Innsbruck Model of Foreign Language Teaching and Learning (IMoF). She is involved in numerous in-service teacher training programs nationally and internationally.

Welcome Reception

Welcome Reception

Applied Language Centre

Thursday, May 30, 19.00-21.00

Sponsored by:

BRITISH COUNCIL

ASSESSMENT RESEARCH GROUP

Conference Programme Overview

Friday, May 31, 2019

08.00-09.00	**Registration**	Foyer, George Moore Auditorium
09.00-09.30	**Opening Ceremony**	George Moore Auditorium
09.30-10.30	**Keynote 1**	George Moore Auditorium

Transitions in language assessment: Topics en vogue, in the doghouse, inert despite lip service, and back from the dead

Talia Isaacs
Chair: Barry O'Sullivan

10.30-10.40 **Festschrift for Sauli Takala**

Claudia Harsch, Neus Figueras, Ari Huhta, Gudrun Erickson

10.30-11.00 **Coffee**

11.00-12.30 **Parallel Papers**

Chair: Eli Moe Auditorium

Automated assessment of fluency and pronunciation in spontaneous speech: Implications for automated speech scoring
Ching-Ni Hsieh

Assessment of L2 English oral proficiency – didactic transpositions from policy documents to operationalization
Liliann Byman Frisén

The CEFR companion volume and rating scale revision: A case study focusing on the updated phonological control scale
Philip Horne

Chair: Alex Runchman E0.01

Challenging paradigms: The role of test taker subjectivities and agency in shaping the consequences and validity of testing within Australia's skilled migration policy
Kellie Frost

The impact of EPLIS on aviation English teachers
Paula Souza

Online diagnostic testing for young learners – An international trial
Tony Clark and Heidi Endres

12.30-14.00 **Lunch** (not provided – choice of restaurants on campus)

12.30-13.30 **Lunchtime Sessions** Applied Language Centre

Room 3	Room 2	Room 15	Self Access
Getting Published Luke Harding, Paula Winke & Constant Leung	**Writing an Abstract** Barry O Sullivan & Talia Isaacs	**Giving a Good Presentation** Bart Deygers & Sonja Zimmermann	**Newcomers to EALTA** Jamie Dunlea & Neus Figueras

14.00-15.30 **Parallel Papers**

Chair: Doris Frötscher Auditorium	*Chair: Norman Verhelst* E0.01	*Chair: Neus Figueras* H1.51
Reading in a new technology environment: Are reading assessments still in the ballpark? *Monique Reichert and Charlotte Krämer*	**Technology and authenticity in the IELTS Speaking Test: What do examiners say?** *Nahal Khabbazbashi, Daniel Lam and Fumiyo Nakatsuhara*	**CEFR practices and technology-based language assessment in higher education** *Karen Ní Chlochasaigh and Tj Ó Ceallaigh*
Double play in listening assessment: Towards increased authenticity *Franz Holzknecht*	**Exploring yes/no questions to promote authenticity in L2 oral performance assessments** *Veronika Timpe-Laughlin*	**Language learning outcomes at the end of basic education and upper secondary education in Finland** *Raili Hilden and Juhani Rautopuro*
Investigating test method effects in French L2 reading items for young learners *Peter Lenz, Katharina Karges and Malgorzata Barras*	**TBLA in an online environment. Opportunities and challenges for authenticity and learner autonomy?** *Goedele Vandommele*	**What happens if L1-students take a high stakes B2-L2-test? Some unexpected results** *Elke Gilin and Lieve De Wachter*

15.30-16.00 **Coffee**

16.00-17.30 **AGM** George Moore Auditorium

19.00-late **Social event** – Pub visit (Dublin city centre)

Saturday, June 1, 2019

09.30-10.30 **Keynote 2** George Moore Auditorium

Computational linguistic analysis, assessment, and language development: Considering language and task

Detmar Meurers

Chair: Peter Lenz

10.30-11.00 **Coffee**

11.00-12.30 **Parallel Papers**

Chair: Carolyn Westbrook Auditorium	*Chair: Benjamin Kremmel* E0.01	*Chair: Ari Huhta* H1.51
Academic language proficiency as a predictor of achievement of first-year university students *Jordi Heeren*	**Transitions in L2 writing assessment – Insights from eye tracking and stimulated recall** *Sonja Zimmermann*	**An investigation into assessing the pragmatic competence of ESL learners at B2-C2 levels** *Edit Willcox-Ficzere*
Language tests for student admission: Transitions to EMI in higher education *Slobodanka Dimova*	**The use of eye-tracking and verbal protocols in construct validation: Multiple-text reading tasks in EAP tests** *Aylin Unaldi*	**Comparing writing skills in different languages using the same scale** *Louise Courtney*
International students enter German university: An empirical study of language proficiency and academic success *Katrin Wisniewski and Jupp Möhring*	**Using PRAAT to measure fluency construct in TEEP speaking tests** *Parvaneh Tavakoli, John Slaght and Gill Kendon*	**Researching academic reading in two contrasting English-medium university contexts and implications for the design of TOEFL iBT** *Nathaniel Owen, Prithvi Shrestha and Kristina Hultgren*

12.30-14.00 **Lunch** Main Restaurant, UCD

Sponsored by Trinity College London

Poster Sessions & Networking Opportunities

Foyer, O’Brien Centre

Developing a list of empirical English word difficulties
Steven Lattanzio and Alistair Van Moere

From traditional to online: Standardisation trainings for oral examiners of different languages
Zoltán Kiszely

A comparative study of in-service language teachers’ beliefs on assessment
Dina Tsagari and Karin Vogt

Towards increased authenticity in integrated writing tasks: Construct operationalization in rating scales
Santi Lestari

Assessing low-level writing: identifying the need for change
Veronika Schwarz, Franz Holzknecht, Eva Konrad and Carol Spöttl

Striving for authenticity in testing listening
Doris Froetscher and Nikolaus Giffinger

Large-scale test accommodations: from practicality to a research and validation agenda
Richard Spiby and Judith Fairbairn

Conceptual and practical challenges in assessing young learners’ foreign language skills
Ari Huhta and Karoliina Inha

Classroom-based Assessment of Oral Mediation: Challenges and Opportunities
Olga Lankina and Yulia Pets

Finding equal balance between standardised tests and classroom-assessment
Yevgeniya Pronoza

Reading for success: investigating readers’ cognitive processes in Austrian EFL reading tests
Klaus Siller and Andrea Kulmhofer-Bommer

Scientific conference: an authentic environment to assess medical students' English communication skills
Eva Braidwood and Magdalini Liontou

Language Assessment Literacy in a Saudi Context
Arwa Alyami

14.00-15.15 Parallel Works-in-Progress

Chair: Stergiani Kostopoulou Auditorium	*Chair: Dina Tsagari* E0.01	*Chair: Richard Spiby* H1.51
Developing an L2 speaking test corpus: Construction and analysis of a pilot corpus *Luke Harding, Dana Gablasova, Vaclav Brezina, John Pill and Jamie Dunlea*	**From language class to higher education: assessing refugees in transition** *Anika Müller-Karabil and Claudia Harsch*	**Technology-assisted scoring of short-answer items for listening comprehension: A clustering approach** *Leska Schwarz and Christian Gold*
Assessing teacher discourse in a spoken English proficiency test *Odette Vassallo, Daniel Xerri and Larissa Jonk*	**German high school students‘ preparedness for English-medium Bachelor programs** *Christine Ringwald*	**Instruments to evaluate the development of intercultural competence** *Maria Leticia Temoltzin*
Diagnostic-dynamic assessment - conceptual and practical innovation in foreign language assessment *Matthew Poehner, Ari Huhta and Dmitri Leontjev*	**Use of feedback practices at language classes in Finnish upper secondary schools** *Toni Mäkipää*	**Change in policy - change in practice?** *Doreen Spiteri*

15.15-16.00 Coffee

16.00-17.30 Symposia (parallel) George Moore Auditorium/E0.01

Chair: Zoltán Lukacsi Auditorium	*Chair: Karin Vogt* E0.01
Language tests for citizenship purposes and low-literate learners *Lorenzo Rocca, Jascha Rüsseler, Bart Deygers, Cecilie Hamnes Carlsen* **Discussants:** *Kellie Frost, Susy Macqueen*	**Transitions in the use of C-tests** *Meg Malone, Yiran Xu, Benjamin Kremmel, Steffen Brandt, Merve Demiralp* **Discussant:** *Claudia Harsch*

19.30- 22.30

Fallon & Byrne, Exchequer Street

BRITISH COUNCIL

ASSESSMENT RESEARCH GROUP

SUPPORTING RESEARCH AND LEARNING IN ASSESSMENT

The British Council supports research and learning in assessment in a number of ways. The Assessment Research Group offers research grants, provides free learning materials, publishes reports and undertakes research projects. You can access free training videos, materials and publications from the website links below.

Offering awards and grants

The British Council offers a range of research awards and grants for research in language assessment.

Assessment Research Awards and Grants: These awards and grants recognise achievement and innovation within the field of language assessment. They are aimed at both research students and more experienced researchers in the area of assessment. See: **www.britishcouncil.org/exam/aptis/research/assessment-advisory-board/awards**

Reading into Research Grants: MetaMetrics and the British Council Assessment Research Group invite applications for research which will contribute to our understanding of the construct of EFL reading comprehension and reading comprehension assessment. Grants are offered to qualifying institutions and/or individuals. **www.britishcouncil.org/research-reading-grants-scheme**

Providing research publications

We publish research across a range of assessment topics and areas of expertise.

- **Technical Reports:** primarily focused on the test development and validation studies related to the Aptis test system.
- **Assessment Research Awards and Grants Reports:** projects carried out by external researchers that have been funded through our awards scheme.
- **Non-Technical Summaries of ARAG Reports:** short, non-technical overviews of the Assessment Research Awards and Grants Reports listed above.
- **British Council Validation Series:** studies in collaboration with external researchers to target areas of importance for Aptis and for language assessment generally.

These publications can be downloaded free at: **www.britishcouncil.org/exam/aptis/research/publications**

Collaborative research with China

In 2016, the Ministries of Education of the UK and China agreed to conduct collaborative research on linking various English language tests to China's Standards of English Language Ability. The National Education Examinations Authority (NEEA), Ministry of Education, China and the British Council were appointed to implement this joint programme. The two-year project brought together experts from the British Council's Assessment Research Group, Cambridge Assessment English, NEEA and leading Chinese universities to develop a model of best practice for linking examinations to the CSE. To find out more about the study, see: **www.britishcouncil.cn/en/exams/cse/results**

Training: free videos and glossary

How Language Assessment Works is a project providing free information, materials and training on language assessment. Our short, animated videos give you an insight into some of the main topics. The practical skills topics have accompanying worksheets and answer keys.

We have recently published a Glossary, consisting of hundreds of definitions of terms related to language assessment. Experienced practitioners wrote the definitions, with language teachers in mind.

Watch the videos or download a free copy of the Glossary at: **www.britishcouncil.org/exam/aptis/research/assessment-literacy**

Evaluating English language capability

The English Impact projects 2017 surveyed the English language capability of a representative sample of 15-year-old students in four countries: Sri Lanka, Bangladesh, Spain (Madrid) and Colombia (Bogota). Capability looks at both current achievement and future opportunity to succeed. The projects provided a comparable baseline showing levels of English language capability in publicly-funded schools, which is where government policy makes an impact.

To find out more or to download the reports, see: **www.britishcouncil.org/exam/aptis/research/english-impact**

www.britishcouncil.org/exam/aptis/research

Sunday, June 2, 2019

09.30-11.00 **Keynote Symposium** George Moore Auditorium

Chair: Jamie Dunlea

Validity Models in Transition
Michael T. Kane, Talia Isaacs, Detmar Meurers, Barry O'Sullivan, Claudia Harsch, Constant Leung **Discussant:** *Luke Harding*

11.00-11.30 **Coffee**

11.30-12.30 **Keynote 3** George Moore Auditorium

Flexibility and utility in assessment and validation

Michael T. Kane
Chair: Claudia Harsch

12.30-13.00 **Closing ceremony** George Moore Auditorium

13.30-15.30 **Guided tour**

Guinness Brewery, St James's Gate, Dublin

Friday, May 31, 2019

08.00-09.00 **Registration** Foyer, George Moore Auditorium

09.00-09.30 **Opening Ceremony** George Moore Auditorium

Dr. Claudia Harsch, EALTA President

Professor Dolores O'Riordan, Vice-President for Global Engagement, University College Dublin

09.30-10.30 **Keynote 1**

Transitions in language assessment: Topics en vogue, in the doghouse, inert despite lip service, and back from the dead
Talia Isaacs

In line with the theme of this year's EALTA conference on transitions in language assessment, this keynote will overview and unpack selected terms, themes, trends, and topics in our field, charting how community perceptions of domain relevance, social appropriateness, and importance to stakeholders often shift over time though sometimes experience stasis. The discussion will first center on transitions in fundamental aspects that define who we are and what we do, including how we refer to ourselves as a field (e.g., language testing; language assessment; language testing and assessment), the move away from a sole focus on psychometrics to the integration of qualitative and mixed methods research paradigms in dissemination, growth in the number of active researchers in our community with greater gender balance and diversity, including through the addition of new regional associations, and evolutions in debates about what the mission of our professional associations should be. Substantive issues that will be covered will relate to shifts in attitudes on the use of technology in administering and scoring tests and the problem of defining a suitable standard for high level performance, with examples from pronunciation research, which has gone from being at its peak to off the radar and now reborn. The talk will then present research on the inclusion of linguistic minority patients in clinical trials that has been perceived on the one hand as being highly innovative in advancing our field while, on the other, as drawing too strongly from another discipline to be germane to our community. Taken together, these examples will demonstrate that the notion of whether or not something is in or out of the language testing/assessment box is clearly in the eye of the beholder.

10.30-10.40 **Festschrift for Sauli Takala**
Claudia Harsch, Neus Figueras, Ari Huhta, Gudrun Erickson

10.40-11.00 **Coffee**

11.00-12.30 **Parallel Papers**

George Moore Auditorium
11.00-11.25

Automated assessment of fluency and pronunciation in spontaneous speech: Implications for automated speech scoring

Ching-Ni Hsieh

The field of language testing has seen a growing popularity in the use of automated scoring in large-scale language assessments (Evanini, 2018). Automated scoring, when used, is often restricted to tasks that are highly predictable, such as read aloud. Recent advancements in spoken language processing technology have led to significant improvements in the performance of automated scoring of open-ended spoken responses. In this interdisciplinary study, we explored components of spontaneous speech delivery as conceptualized in second language acquisition and language testing research (Ginther, Dimova, & Yang, 2010; Isaacs, 2014) and investigated the relationships between human judgments of speaking proficiency and machine generated features of fluency and pronunciation of adult L2 speakers. Drawing on more than 200,000 speech samples collected from TOEFL iBT test takers who responded to six open-ended speaking items, we analyzed the empirical relationships between human scores on the spoken responses and an automated speech scoring system, SpeechRater, generated fluency and pronunciation features of each response. The analyses yielded moderately high correlations (|r | > 0.4) for pause frequency, speaking rate, mean length of run, segmental pronunciation, and word stress patterns. Weak correlations, in contrast, were found for repair fluency (e.g., repetitions, self-corrections), and a few prosodic features. The differential magnitudes of correlations partially resulted from the machine's ability to accurately identify repair phenomena and to mark intonational phrases in spontaneous speech. Implications for automated speech scoring and how to enhance assessment literacy in this area will be discussed (Harding, 2018).

11.30-11.55

Assessment of L2 English oral proficiency – didactic transpositions from policy documents to operationalization

Liliann Byman Frisén

In Sweden, students in grades 6 and 9 take a national high-stakes oral proficiency (OP) test in L2 English, the National English Speaking Test (NEST). Teachers are to assess NEST holistically guided by overarching descriptors. When operationalizing assessment of OP, many teachers-as-raters construct their own scoring rubrics where they list what to focus on in the assessment (cf. Bøhn, 2015). The aim of this study is to gain new knowledge about how teachers-as-raters interpret and transpose the test construct – 'oral production and interaction' – as reflected in their own scoring rubrics. Data consist of 20 scoring rubrics made by teachers (N = 17) for the assessment of the NEST (grades 6 and 9). For demographic data about the teachers, a web questionnaire was used. Rubrics as well as questionnaires were collected through two closed Facebook groups for L2 English teachers in Sweden. Scoring rubrics data were analyzed qualitatively, adopting content analysis guided by construct, criteria and sub-criteria (Bøhn, 2015), to examine how aspects of OP were used for assessment and structured by the teachers, as well as how these aspects were interpreted and transposed (Achiam & Marandino, 2014; the Anthropological Theory of Didactics, Chevallard, 2007). Since the scoring rubrics were created by the teachers themselves they provide insight into L2 teachers' practices as well as their conceptualizations of assessment, topics that have been shown to be under-researched (Brookhart, 2018; Roca-Varela & Palacios, 2013). The study promotes discussions on national assessment guidelines, operationalization of OP test constructs, and inter-rater reliability.

12.00-12.25

The CEFR companion volume and rating scale revision: A case study focusing on the updated phonological control scale

Philip Horne

The CEFR Companion Volume (CV) presents a transition point in the benchmarking of tests against the CEFR. This is especially important in this case study, because the test in question is not only retrospectively aligned against the CEFR, but tasks and rating scales are based upon it a priori. This transition, therefore, calls for re-examination of the benchmarking. The presentation reports on a study focusing on the updated Phonological Control Scale. This has clear implications for the CEFR-linked pronunciation descriptors in the holistic rating scale utilised in the case study. There are also broader issues concerning pronunciation descriptors within holistic rating more generally. Research questions focused, therefore, on the nature of pronunciation-related decision-making in holistic rating, and the extent to which a more detailed operationalisation of pronunciation descriptions (based upon the CV) would refine scoring.

The study took a mixed-methods approach. 6 raters rated a number of speaking tasks (n=50), using both the existing rating scales and the updated Phonological Control Scale. This was followed by a series of paired interviews. Results indicate that the current pronunciation descriptors lack sufficient detail, in particular at higher levels. There was high correlation between the two sessions (rs = .91), indicating that pronunciation can factor more greatly into decision-making in holistic rating. Furthermore, a chi-squared test revealed that pronunciation-based decisions occur less frequently at higher levels, indicating the descriptors in their current format are less relevant to rater decision-making. The session concludes by suggesting modifications based upon the CV's updated Phonological Control Scale.

E0.01

11.00-11.25

Challenging paradigms: The role of test taker subjectivities and agency in shaping the consequences and validity of testing within Australia's skilled migration policy

Kellie Frost

Within Australia's skilled migration policy, English test scores play a central and increasingly restrictive role in migrant selection processes and the allocation of permanent residency rights. This paper reports on a study of the English test experiences of 22 international accounting graduates seeking permanent residency as skilled migrants in Australia. An interview-based, grounded theory approach focused on identifying how these graduates perceived the score requirements for permanent residency, their perceptions of different tests, and how these perceptions shaped actions as individuals struggled to realise their migration intentions. Findings demonstrate that testing in this policy space undermined graduates' engagement with work and hindered the development of work-ready English skills, thereby subverting skilled migration policy aims. Moreover, findings revealed that wide networks of information sharing among test takers play a key role in shaping interpretations of test constructs and approaches to test preparation and test taking, generating validity issues relevant to particular tests and to the practice of testing more broadly within this policy space. Our study highlights a lack of alignment between the ways tests actually function in this policy domain and the notion of test purpose as imagined by language testers, and foregrounds a need to re-situate the experiences of those subjected to testing regimes, rather than the intentions of test developers, as the nexus of evaluations of validity and fairness in high stakes immigration policy contexts.

11.30-11.55

The impact of EPLIS on aviation English teachers

Paula Souza

This study investigated the impact of EPLIS (Aviation English Language Proficiency Exam for the Brazilian Airspace Control System) on teachers' perceptions, attitudes and actions in an Air Traffic Control Initial Training Program. EPLIS was developed in 2007 to comply with ICAO's (International Civil Aviation Organization) language proficiency requirements for professionals involved in international flight operations. For the first 7 years, EPLIS was delivered only to in-service air traffic controllers, but in 2014, its application was extended to initial training programs to supposedly promote curricular changes aiming at improving learners' proficiency. However, the consequences of this decision, whether intended, unintended, positive or negative, have not been appraised yet. Firstly, a 65-item questionnaire was designed and delivered to a group of 21 teachers. 16 of them answered it and the data were statistically analyzed for patterns and trends of perception and behavior among the participants. Then, semi-structured interviews were conducted with 4 voluntary teachers who also had their classes observed during a period of 3 weeks. The results showed that the decision to introduce EPLIS in the school context actually increased its impact. However, deficiencies in the understanding of the exam and its demands were still perceived among some teachers, compromising the intended effects. The findings revealed that in order to promote changes, besides offering teachers training in the test, it is necessary to develop their language assessment literacy mainly because they are the ones responsible for constructing the classroom-based tests, which were considered higher stakes than EPLIS.

12.00-12.25

Online diagnostic testing for young learners – An international trial
Tony Clark and Heidi Endres

An effective diagnostic test can play a key role in students' transition from struggling EFL learner to effective language user; allowing specific strengths and weaknesses in linguistic development to be identified prior to instruction (Jang, 2012). This paper describes an online diagnostic test developed by Cambridge English that assesses receptive English grammatical knowledge at A2 level, in order to help students make this transition.

Aimed at young learners (approximately 15 years old), the test provides detailed diagnostic feedback on seven grammatical categories at both individual and class levels, helping improve the effectiveness of curriculum planning and to accommodate students' learning needs. Specifically, this instant automated feedback informs teachers which aspects of grammar to focus on to reach A2 level and suggests suitable teaching and learning materials to achieve this. The test was trialled in Estonia, Mexico, Peru and Spain, and using data from surveys (1,025 responses) and seven focus groups, student and teacher perspectives on the test were analysed. Emerging themes included consideration of local educational contexts, validity of automated feedback, implications of computer-based delivery, practical challenges of test implementation and usefulness of feedback. Participants were largely positive about the experience, but had recommendations for improvement – particularly regarding delivery of feedback and expansion to include other proficiency levels and skills. Based on these perspectives, a series of modifications for the next test iteration are outlined. Finally, the implications of this research for wider pedagogical practices are considered, and further suggestions to improve the transition towards targeted learning made.

12.30-14.00 **Lunch**

14.00-15.30 **Parallel Papers**

George Moore Auditorium

14.00-14.25

Reading in a new technology environment: Are reading assessments still in the ballpark?

Monique Reichert and Charlotte Krämer

Using digital technologies for daily activities such as communicating or learning has become ubiquitous today. This trend is also clearly visible within the field of reading habits, especially among adolescents: Recent studies underline that traditional text types (e.g., fiction books) are no longer part of the more commonly read text materials (Duncan et al., 2016). The question addressed in the presentation will deal with the degree to which different reading habits impact on adolescents' reading competence, and is intended to encourage a discussion about the construct of reading competence as generally operationalized in reading competence assessments. We base our considerations on the analyses of two data sets: The first one deals with the extra-curricular reading habits of 3074 9th grade students, and the impact of these practices on their German reading competence. The corresponding data are taken from a survey regarding their reading habits in terms of ten different types of texts (e.g., non-fiction books, e-mails). A regression analysis reveals the strongest impact on reading competence for narrative texts, whilst reading digital texts – although highly attractive among the students – is found to be of minor importance. The second data set differentiates between the reading habits of around 4500 adolescents regarding 1) traditional (printed) texts, 2) digitalised texts (e.g., e-books), and 3) text types that have emerged with new technologies (e.g., social media texts). The ensuing discussion compares particularities of classical and digital texts, and raises questions concerning the construct of reading competence (to be) targeted by standardized tests.

14.30-14.55

Double play in listening assessment: Towards increased authenticity

Franz Holzknecht

All listening test developers are faced with a choice: whether to play the listening text once (single play), or twice (double play). Previous research has investigated the effects of single and double play on test scores, however an equally relevant question is whether repeating the listening text impacts construct validity (Field, 2015). It has not been fully established in what ways double play influences candidates' cognitive behaviour and thus how it alters the coverage of the listening construct.

This study investigated the effects of double play on test scores and candidates' cognitive behaviour. In the first stage, 316 candidates responded to two multiple choice and two open format listening tasks in single and double play and answered questionnaires targeting strategic behaviour and anxiety levels. Data were analysed through Rasch analysis and inferential statistics. In the second part, 16 candidates completed the same tasks in both conditions on an eye-tracker and performed verbal recalls, which were stimulated by their eye-movements while they had been solving the items. Recalls were coded according to candidates' strategic behaviour, anxiety levels, and cognitive processes. The results indicate that double play yielded higher scores, supporting previous findings. However, it also had benefits for construct validity. Specifically, candidates used fewer test-taking strategies, were less anxious, and displayed a greater variety of listening strategies and cognitive processes compared to single play. Implications for test development are discussed, in particular the role of double play in enhancing authenticity in language assessment.

15.00-15.25

Investigating test method effects in French L2 reading items for young learners

Peter Lenz, Katharina Karges and Malgorzata Barras

In large-scale assessments, selected-response item formats, such as multiple choice (MC) are often preferred over open-ended (OE) formats because they are cheaper to score. But do they measure the same construct? – In his meta-analysis on item types, Rodríguez (2003) found that MC items are generally easier but do not represent a different construct, especially stem-equivalent items. Ozuru and colleagues (2007; 2013), however, identified reading task conditions in which the two types performed differently.

In our study involving roughly 600 learners of French at the end of primary school, we explored the behavior of stem-equivalent MC and OE reading items. To focus on differences between the constructs embodied by the two item types, we used a 2-dimensional Rasch model to construct for each pupil a measure based on MC items and another one based on OE items. We then used the results from a number of precursor skills tests (e.g. vocabulary) and integrative tests (e.g. C-Test), taken by the same students, to predict success on the MC and the OE dimensions (using repeated-measures ANCOVA). The main findings are as follows: correlation between both dimensions is high enough to assume rough construct equivalence. Also, receptive vocabulary knowledge (Yes/No test) is the best predictor of success on both item types – but significantly more so on the MC-based items. In the complete model, differences in working memory have a significant influence on reading performances based on OE items only, which may indicate that these involve more active language processing.

E0.01

14.00-14.25

Technology and authenticity in the IELTS Speaking Test: What do examiners say?

Nahal Khabbazbashi, Daniel Lam and Fumiyo Nakatsuhara

As the world of language testing transitions towards embracing technology in the digital era, new questions and debates emerge about authenticity and construct representation, particularly in relation to speaking assessment.

Our IELTS-funded project draws on examiner voices, widely held as a valuable source for informing, refining, and revising speaking test construct(s), to explore the themes of technology and authenticity in the context of the IELTS speaking test. Over 1,200 IELTS Speaking Test examiners worldwide participated in this large-scale survey bringing together insights about the current format of the test and possible improvements in the future.

Using a sequential explanatory design, the first quantitative stage of the study involved the administration of an online survey to examiners on various aspects of the speaking test such as tasks, topics, format, interlocutor frame, training and standardisation procedures. The second qualitative stage involved in-depth exploration of key issues identified from the survey through semi-structured interviews with a representative sample of 35 respondents.

The findings revealed a tension where, while 95% of the survey respondents expressed a preference for face-to-face over computer-based mode for delivering the IELTS Speaking Test, the examiners interviewed voiced concerns over the authenticity of spoken interaction elicited in the current test, which has a highly structured interlocutor frame. Examiners also offered insights on ways to improve the authenticity of current speaking test topics, moving away from dichotomous or reductionist categorisations of gender, background, and culture that are so deeply ingrained in assessment practices.

14.30-14.55

Exploring yes/no questions to promote authenticity in L2 oral performance assessments

Veronika Timpe-Laughlin and Innhwa Park

Yes/no questions have been found to feature prominently in naturally occurring talk-in-interaction (Raymond, 2003; Stivers, 2010). However, in second language (L2) assessment, there is a concern that yes/no interrogatives only allow limited answers and thus restrict L2 learners' performance. This discrepancy poses a dilemma for L2 test designers who aim for 'authenticity' (i.e., approximation or replication of language use in a real life) in their oral performance assessments.

To address this mismatch, this study focused on the extent to which yes/no questions in an L2 role-play assessment task elicited language reflective of real-life communication. The study was guided by the following research questions:

(1) How do L2 English learners respond to a yes/no interrogative in a role-play assessment context?

(2) To what extent do learners' responses align with patterns found in naturally occurring talk-in-interaction?

We used conversation analysis to examine how L2 learners respond to yes/no questions in 104 role-play conversations between an L1 American English speaker and L2 English learners. We found that yes/no interrogatives elicited a range of responses (e.g., repair initiation, type-conforming responses, non-type-conforming responses) that also resulted in negotiation of meaning between participants. While the turn design and sequence organization features we observed largely align with the existing literature on naturally-occurring conversations, we will also discuss deviant cases that provide insights into L2 learner discourse. Finally, we discuss the findings in terms of implications for L2 assessment.

15.00-15.25

TBLA in an online environment. Opportunities and challenges for authenticity and learner autonomy?
Goedele Vandommele

Task-based language assessment (TBLA) strives to assess appropriate and effective language use via tasks ‘that reflect the tasks and interactions that learners are expected to perform in real-life situations, within a particular domain’ (Van Gorp & Deygers, 2013). Authentic tasks and their development, therefore, are at the core of TBLA. As the world is changing, so is authentic language use: in the 21st century, real-life communication often happens digitally. TBLA in a digital environment therefore might (more) accurately reflect real-life communication.

Following this line of thought, this paper focusses on the opportunities and limitations offered for task-based assessment by the transition from a paper-and-pen to a digital test. Central to the discussion is The Certificate of Dutch as a Foreign Language (CNaVT), a task-based standardized test in transition, and the data gathered and analyzed in the design and trialing phases of the new online test construction.

In this paper, we will present (a) different online tasks in a design and trial phase and (b) quantitative (survey) and qualitative (focus group conversation) information on the reception of these tasks by the various stakeholders of the CNaVT, all the while focusing on how online testing can reaffirm and/or challenge the traditional strengths of TBLA: does augmenting situational and interactional authenticity provides high content and face validity (Ross, 2012)? Moreover, possibilities and limitations of online testing for increasing learner autonomy are considered.

H1.51
14.00-14.25

CEFR practices and technology-based language assessment in higher education
Karen Ní Chlochasaigh and TJ Ó Ceallaigh

The authors deliver a blending learning programme for Irish language immersion educators, through interactive onsite workshops and online activities, self-directed tasks and synchronous workshops. Participants gain entry at level B1 of the CEFR and develop their language skills towards level C1 through participation and the completion of specific target level tasks to assess and gain proficiency. A cohort of 23 students completed an online questionnaire detailing their specific linguistic and pedagogical needs, their personal goals and recommendations. Questionnaire data was used to carry out a needs analysis upon which content themes were identified to inform the design of assessment methods in a relevant and meaningful approach. Technology-based language assessment forms a core element where a model of self-reflective, collaborative and professional practice assessment is implemented. Students receive corrective and informative feedback on assessments which correspond directly to a grading rubric and over time build an e-learning portfolio, creating the opportunity to inform pedagogical practice and policy and practice in leadership. This paper would share experiences, gained knowledge and developed best practices in relation to transitioning towards technology-based language assessment, using examples of assessment design and data to highlight changes in the assessment landscape and its effects on language teaching and learning. Chapelle (2003) called for 'much theoretical and empirical work to bridge the gap from current technological capabilities to progress in language assessment'. This paper aims to discuss and evaluate some of those innovations and benefits in technology-based language assessment in relation to a specific target level of proficiency.

14.30-14.55

Language learning outcomes at the end of basic education and upper secondary education in Finland
Raili Hilden and Juhani Rautopuro

The study is a longitudinal comparison of students (n=1491) language ability at the end of compulsory education and upper secondary school leaving exam. It sheds light on the transition between educational stages and the impact of upper secondary education. Learning outcomes in English and Swedish were investigated in relation to certain background variables at both points of time.

> RQ1. What kind of associations are detected between the level of language proficiency in English and Swedish at the end of basic education compared with the level at the end of upper secondary education?
>
> RQ2. What group-wise differences are detected between certain background variables (e.g. gender, parents´ educational level, plans for future studies) in both languages?

The linguistic test items were set onto levels of The Common European Framework scale, applying a standard setting procedure for listening and reading. Five-step Likert scale was used to grade perceptions, and the data as a whole were subjected to standard analyses of inferential statistics (correlative and regression analyses).

The results of this study corroborate the validity of the national evaluation of learning outcomes at the end of basic education as predictors of success in upper secondary studies. However, it also witnesses about persistent group-wise differences in the outcomes at both points of time, which can be considered as a challenge with regard to equality, and needs ongoing attention in educational policy making. We invite the participants to discuss these issues.

15.00-15.25

What happens if L1-students take a high stakes B2-L2-test? Some unexpected results

Elke Gilin and Lieve De Wachter

High stakes university entrance tests for L2-speakers are assumed to measure the language proficiency needed for academic success (Deygers and Carlsen 2014; 2015, Deygers et al. 2017). Although little research has supported the claim that B2 is the level needed to cope with the linguistic demands of university education (Deygers 2017), even fewer studies have investigated the claim that L1-speakers automatically have this B2-level.

In Flanders, Belgium, there are no entry requirements for L1-speakers, except for the faculty of Medicine. Consequently, the influx of students in higher education is large, many of them not passing their first year. In that respect, it is interesting to investigate how Flemish secondary school students entering higher education without a test score on a language test designed for L2-students, measuring the B2-level.

This study focuses on the results of a small-scale study carried out in Flanders with last year secondary school pupils (N=50). All pupils took the computer test of the Interuniversity Test Dutch as a Foreign Language (ITNA) testing reading, listening and language in use on a B2-level. The test has an ALTE Q-label.

It is interesting to see that 32% of the students didn't pass the threshold. Especially pupils with a multilingual background seemed to have difficulties. This not only challenges the justice and fairness of the test for L2-students, but it also provokes discussion about the open entrance system for L1-students. During our presentation we will make suggestions on redesigning the entrance policy for higher education in Flanders.

15.30-16.00 **Coffee**

16.00-17.30 **AGM** George Moore Auditorium

Saturday, June 1, 2019

09.30-10.30 **Keynote 2** George Moore Auditorium

Computational linguistic analysis, assessment, and language development: Considering language and task
Detmar Meurers

Approaches automating the analysis of language to support formative or summative assessment analyze different aspects of language and rely on different task-specific information. In automatic essay scoring (Burstein 2013), we find both: a) topic-specific models that compare student essays to large sets of manually scored essays written for the same prompt - based on the analysis of few aspects of language, such as latent semantic spaces derived from the essay as a "bag of words", and b) analyses integrating a wider range of linguistic analyses and writing constructs. For short answer assessment (Ziai, Ott, Meurers 2012), e.g., evaluating answers to reading comprehension questions, the approaches also make use of reference answers (or more explicit scoring rubrics) for the specific task, but given the limited amount of language material available in a single answer and the limited amount of labeled training data available in most settings, the methods generally compare student and reference answers using a broader range of linguistic representations, including abstractions such as lemmas, groupings such as dependency triples or phrasal units, and aspects of the context, such as the question determining the information structure of the answer (Ziai & Meurers 2018).

When it comes to characterizing language development, a substantial strand in Second Language Acquisition research explores the development of linguistic complexity in terms of a broad range of language characteristics (Housen et al. 2019). We illustrate in Alexopoulou et al (2017) that the development of linguistic complexity that so clearly emerges in automated linguistic complexity analysis of large learner corpora can quickly disappear when zooming in on specific writing tasks and the substantial task effects they produce. This brings some focus back from general language analysis to the tasks and the need to make their characteristics explicit in relation to the language production they support (Quixal & Meurers 2016). As Wisniewski's (2017) points out, this also means that including a broader range of tasks in learner corpus work will be important to empirically enrich CEFR research.

An interesting opportunity for directly linking task properties and learner language properties arises when applying the automated analysis of linguistic complexity to continuation writing tasks (Wang & Wang 2015). We show in Chen & Meurers (2019) that in such a setting one can observe alignment for a broad range of linguistic complexity characteristics of the language produced by the learner to the language read. On the practical side, this makes it possible to explore fostering language acquisition by providing input in the "Zone of Proximal Development" that learners align with.

In this talk, we aim to color in some of this sketch of the landscape of computational linguistic analysis in the context of (language) assessment.

10.30-11.00 Coffee

11.00-12.30 Parallel Papers

George Moore Auditorium

11.00-11.25

Academic language proficiency as a predictor of achievement of first-year university students

Jordi Heeren

Starting a university study is an important educational transition point in which language proficiency is assumed to play a role. In Flanders, where, except for Dentistry and Medicine, there are no entry requirements, the influx of students is large and diverse. By consequence, a post-admission language assessment to screen all starting students, including L1-speakers, was implemented in several faculties. In our study we investigated academic language proficiency (ALP) as a predictor of achievement.

The screening is web-based, low-stakes and consists of mostly selected-response, academic vocabulary and reading items. Our construct of ALP resonates with Hulstijn's (2015) concept of Higher Language Cognition (HLC). He claims that L1-speakers' differences in performance on HLC-tasks will be relatively large, due to differences in literacy, cognitive abilities and educational levels.

We correlated the ALP-score and credit completion rate of 12,435 first-year students. In addition, we built a regression model with a subset of 5167 students which controls for students' demographic (gender, age, socioeconomic status, home language and nationality) and educational background (high school average grade and educational track).

Our results show that ALP is a small, but significant predictor of achievement and that it can select a possible at-risk group, proving the usefulness of the instrument in an open university system. The predictive value of ALP is only slightly effected by the demographic variables, but diminishes more clearly when educational background variables are added. This seems to confirm Hulstijn's hypothesis that, in our sample of mostly L1-speakers, ALP partly reflects prior educational achievement.

11.30-11.55

Language tests for student admission: Transitions to EMI in higher education

Slobodanka Dimova

While many studies investigate the validity of TOEFL and IELTS scores, as well as stakeholders' opinions about their usefulness for international student admission at Anglophone universities (Ginther & Elder, 2014; Hyatt, 2013), little is known about their validity and usefulness for admission at non-Anglophone universities that are transitioning to English medium instruction (EMI) (Klaassen & Poot, 2013). Although it is reasonable to expect comparability across university contexts, differences may emerge due to the dissimilar educational traditions, linguistic environments, and contextual conditions. Given the premise that the pragmatics, or the effectiveness, of score interpretation and uses needs to be evaluated within the particular context and for particular populations (Kane, 2018), a closer analysis of English for academic purposes (EAP) test scores in the non-Anglophone contexts is needed. The present mixed-methods study addresses this by examining the pragmatics of English language requirements for university entry in the EMI programs at a large European university. The particular context is operationalized through four contextual dimensions: political/historical, economic, socio-cultural, and academic, and particular populations refers to the EMI student body. Data from a lecturer survey (n=199), interviews (n=16), and documents (policies, reports, websites) are used for the analysis. Results suggest that although lecturers find English proficiency essential for students' academic success in EMI, the different contextual dimensions affect the selection, implementation, and impact of English language requirements, and hence teaching and learning. The discussion will raise awareness about how the implementation of new language requirements, or changes in existing ones, may affect local academic conditions.

12.00-12.25

International students enter German university: An empirical study of language proficiency and academic success
Katrin Wisniewski and Jupp Möhring

In line with higher education policy goals, the number of international students in Germany is continuously growing. However, dropout rates are high (foreigners 41%, Germans 28%; Heublein et al. 2017). While empirical research is scarce, restricted language abilities might play a crucial role. To gain a broader picture, we present a project (2017-2020) that disentangles the relationship between language and academic success while simultaneously assessing the role of other factors of influence such as integration or self-regulation. Language proficiency is measured longitudinally with standardized language tests (e.g., ACTFL LPT and RPT; onSET; TestDaF writing) once per year for three cohorts at two German universities. Academic success is operationalized via credit points, marks, and self-report data based on assessments of study success.

We present first results based on cohort 1 (starting autumn 2017, N=155) and cohort 2 (starting autumn 2018). The results show that although students passed one of the obligatory university admission language tests (B2/C1), 25% of them did not reach B2 in any of our own tests. Furthermore, these students obtained fewer credits in their first semester ($p=.000$, $\eta^2=.345$), displaying substantial relationships between language competence and academic success. Self-report data revealed significant correlations between lower language proficiency and migratory status, study-related problems, L1, and language learning motivation.

These first cross sectional results will be modeled longitudinally, as soon as data are available. Well will discuss possible implications for the current internationalization policy and suggest to introduce obligatory language support not only prior to, but as well after university access.

E0.01

11.00-11.25

Transitions in L2 writing assessment – Insights from eye tracking and stimulated recall

Sonja Zimmermann

Cumming (2013) pointed out that the assessment of academic writing should not solely be based on traditional formats like timed impromptu essays, but also incorporate 'content-responsible' writing. A case in point are integrated tasks that require test-takers to process information from different language-rich source material and integrate this information in their own texts. In the process of revising a large-scale university entrance test for international students in Germany, a summarization task has been added to the writing component. In fact, there has been a change of the test format as well as a change of the test delivery mode (from paper-based to digital).

As part of a larger process-oriented validation study investigating the cognitive processes involved in integrated writing tasks, this paper focuses on the responses and strategies of 19 international L2-learners of German while summarizing from text and graphical input. Following recent approaches (Yu, He & Isaacs, 2017; Révész, Michel & Lee, 2017), the study builds on a mixed-methods design combining eye-tracking and stimulated recall techniques. The eye-tracking data yielded information on (a) reading-writing relations, e.g. the time participants spent in different Areas of Interest (AOIs), and (b) the transitions between the AOIs. In retrospective interviews, test takers commented on various challenging demands of the twofold transition from independent to integrated and from paper-based to digital writing. The paper finally discusses possible washback effects and implications for developing training material.

11.30-11.55

The use of eye-tracking and verbal protocols in construct validation: Multiple-text reading tasks in EAP tests

Aylin Unaldi, Hatice Yurtman Kaçar and Emre Oral

Although reading has long been an extensively researched area, changing dynamics of 21st century educational practices have broadened our understanding of academic reading to include more comprehensive conceptualisations of the skill. A prominent attempt to adapt to these changes is the attempted integration of complex reading tasks in language tests in which information is identified, compared, contrasted and evaluated across multiple information sources. Such reading activities are usual practice especially in higher education (Moore et al. 2010) and are associated with complex reasoning and deeper learning (Cerdan, 2006; Gil et al., 2010).

There are already certain EAP tests that purportedly operationalise multiple-text reading skill. However, there is not any published evidence as to the extent to which these test have attained a successful implementation of the skill. The purpose of this study is to investigate how multiple-text reading skill in tasks in existing English proficiency tests such as ISE II, MET, ECCE, SAT are operationalised, whether the sub-skills and strategies specified in these exams match the theoretical explanations, and whether the actual use of skills and strategies reflects a sufficient and accurate coverage of documents model reading skill. Data were collected from 12 participants through eye-tracking and verbal protocol methods. Preliminary results revealed that certain tasks seem to require a more substantial documents model reading whereas other tasks simply triggered search reading. In this paper, together with the implications of the findings, the useful combination of eye-tracking methodology and think aloud protocols in reading research will be emphasised.

12.00-12.25

Using PRAAT to measure fluency construct in TEEP speaking tests

Parvaneh Tavakoli, John Slaght and Gill Kendon

While assessment of fluency as a key construct of L2 speaking ability in candidates' spoken performance is central to many international tests, little research has been conducted to examine which analytic aspects of fluency distinguish different assessed levels of proficiency. The complexities involved in measuring temporal aspects of fluency objectively has been one of the reasons analytic assessment of fluency has remained a relatively under-researched area. With the development of technology, e.g., PRAAT (Weenik & Boersma, 2013), however, measuring temporal aspects of fluency with a high degree of precision has become possible. In order to help shed light on the relationship between different aspects of fluency and assessed levels of proficiency, we have conducted two studies investigating fluency in Speaking papers of two internationally recognised tests, i.e., British Council's Aptis test and Test of English for Educational Purposes (TEEP). The two data sets came from 90 test takers at A2, B1, B2 and C1 levels of CEFR. Using PRAAT software, the data were analysed for a range of speed, breakdown and repair measures, before being subjected to ANOVAs to examine whether different aspects of fluency were significantly different at different levels of proficiency. The results suggest that although speed measures distinguish between lower levels, they do not distinguish between B2 and C1. While pause position is a helpful measure distinguishing lower from higher levels of proficiency, repair measures do not distinguish levels of proficiency. These findings have significant implications for designing and using rating scales in tests of speaking.

H1.51

11.00-11.25

An investigation into assessing the pragmatic competence of ESL learners at B2-C2 levels

Edit Willcox-Ficzere

As the number of overseas students in English-speaking countries has increased over the last decades, the importance of pragmatic competence in the successful social integration of L2 speakers has been highlighted and the need for assessing it has become greater (e.g. Ross and Kasper, 2013). Most currently available pragmatic tests are based on the Speech Act Theory as a theoretical framework and use discourse completion tasks as test instruments. However, both of these have been criticized lately for overlooking the importance of the discursive side of pragmatics and their inability to elicit more authentic data (e.g. Roever, 2011).

The aim of this research was, therefore, to investigate an approach in assessing B2-C2 level learners' pragmatic competence in extended oral discourse. It aimed to examine the extent to which a dialogic task format allows ESL learners to display their pragmatic competence.

Data were collected from thirty international university students at B2-C2 levels with different L1 backgrounds, who performed two dialogic tasks reflecting authentic situations. This was followed by a semi-structured interview to gain participants' perspectives on the given contexts. Performance of the tasks was video recorded, transcribed and analysed quantitatively as well as qualitatively using a Conversation Analytic framework.

The data from the semi-structured interviews indicated that with increasing proficiency the depth of analysis regarding context also increased, as well as the awareness of the connection between language use and social context. However, only C2 level participants had sufficient cognitive capacity to adjust their language to reflect their own pragmatic intentions.

11.30-11.55

Comparing writing skills in different languages Using the same scale

Louise Courtney

A unique and novel writing assessment has been developed: it is the first in the world to successfully assess writing across languages and scripts in multiple countries with diverse (including lesser taught) languages. The assessment uses common tasks and common scoring criteria, and reports student achievement on a single scale across the multiple countries and languages.

The metric has been developed to survey South East Asian students at a key transition point in their education, the last year of primary schooling, across seven countries in South East Asia.

This paper reports on the framework design, the test development process and the recent results of the Field Trial (15,392 cases) as the main study moves into development in 2019. The socio-political context is described and challenges in developing the language assessments for specific contexts are briefly discussed.

The assessment framework features a literacy orientation, targeting authentic writing life skills, from single word labelling through to freer descriptive writing. Each task is coded using multiple criteria (between two and six criteria, depending on the task). Because of the innovative nature of the Writing component, as a cross-language assessment, it was hypothesised that some features of writing would be assessable across languages whereas others might be language dependent. Field trial data confirming our hypotheses will be discussed.

This heralds a major and exciting transition in language assessment, where practitioners might reconsider the assessment of unrelated languages using a single scale.

12.00-12.25

Researching academic reading in two contrasting English-medium university contexts and implications for the design of TOEFL iBT
Nathaniel Owen, Prithvi Shrestha and Kristina Hultgren

The TOEFL iBT is increasingly used worldwide in the expanding field of English as a medium of instruction (EMI) universities. To date, the TOEFL iBT has been used primarily for admissions decisions for North American universities. However, the extent to which EMI contexts differ from North American, British or other English L1 university contexts remains under-explored. This project contributes to domain analysis of two EMI contexts, Nepal and Sweden, and focuses specifically on an under-researched area in EMI, reading for academic purposes, and assesses the suitability of the reading component of the TOEFL for students entering higher education in those contexts. A mixed-methods study was conducted in Nepal and Sweden in which students (Nepal = 10, Sweden = 9) were asked to complete reading logs over a period of three weeks. These informed the design of a questionnaire (Nepal = 69, Sweden = 60) examining academic reading demands and practices. Students who completed the questionnaire also completed a TOEFL reading test. Follow-up semi-structured interviews (Nepal = 21, Sweden = 23) focused on reading skills and practices. Findings reveal the importance of English reading proficiency to academic success and that significant differences exist in terms of reading demands across the two contexts, indicating the necessity of developing alternative testing solutions for students entering university. Cut scores for traditional English university contexts (e.g. US, UK) may not be appropriate for EMI contexts.

12.30-14.00 **Lunch** Main Restaurant, UCD

12.30-14.00 Poster Session

Developing a list of empirical English word difficulties

Steven Lattanzio and Alistair Van Moere

While there are many lists that categorize words by frequency, this does not necessarily equate to empirical word difficulty. This paper presents the method and results for determining empirical difficulty for individual words using auto-generated cloze items.

The study involved over 38,000 learners who engaged with online reading materials that occasionally clozed out words. In total, 17,966 L1 English students read 110,204 unique passages and encountered 6.7 million cloze items. Further, 20,096 learners of English from 199 different countries read 13,495 unique passages and encountered 3 million cloze items.

It is theorized that when a word is clozed from a sentence, its difficulty can be disambiguated from context when it is presented in many different sentences to many different test-takers. A modified Rasch model was applied to cloze items where the item difficulty is a function of the word being clozed out and text complexity of the surrounding passage.

The results reveal a list of empirical difficulties for 9,000 content words, and also show how empirical difficulties for individual words vary by country (which is a proxy for L1). For example, when comparing word difficulties, Spain and Columbia correlate at r=0.90, while Spain and Vietnam correlate at r=0.74. A correlation matrix for pairings between each top-20 country's shared empirical word difficulties reveals coefficients in the range r=0.7 to 0.9 (e.g. Spain and Columbia, 0.90; Spain and Vietnam, 0.74). The resulting word list can be useful to test content developers.

From traditional to online: Standardisation trainings for oral examiners of different languages

Zoltán Kiszely

The Language Examination Centre of XXX (LEC) provides examinations in five different languages: English, German, French, Italian and Spanish. The centre used to arrange its annual standardisation further training for its oral examiners in a traditional, face-to-face format; however, during the past two years it has been organised online, which consists of two parts. First, the examiners complete a test on their knowledge of the CEFR levels, second, they evaluate an audio recording of an oral language performance. Due to these features of the examination centre the present paper has a threefold aim. The first is to explore whether there are differences between the way examiners of different languages interpret the CEFR scales; the second is to investigate whether there are differences between the way examiners of different languages interpret the criteria of the analytical rating scale used for assessing examinees' oral language performance; while the third is to analyse the difficulties of turning from traditional to online standardisation sessions. The research data in this paper are drawn from the standardisation further training, and an online questionnaire. The participants were 174 English, 80 German, 35 French, 19 Italian and 11 Spanish examiners. The study relates to the conference theme in two ways: First, it provides valuable insights into the considerations of examiners of not only English and German but also those of lesser-taught languages like French, Italian and Spanish, and second, it discusses some lessons learnt on the transition from traditional to technology-based standardisation sessions.

A comparative study of in-service language teachers' beliefs on assessment

Dina Tsagari and Karin Vogt

The field of language teacher education is immense in scope with an evolving research base (Barkhuizen & Borg, 2010), with language teacher cognition being one central area of interest (Borg, 2018). Language teacher beliefs research has focused on areas as diverse as teaching grammar (Schulz, 2001), literacy development (Meijer, et al. 2001) or teacher autonomy (Benson, 2010). Assessment constitutes an important field of teacher activity, given that 30 to 50 per cent of teachers' time is assessment-related (Cheng, 2001). However, although there is a growing body of research on the concept of Language Assessment Literacy (LAL) and the training that English Language Teachers seem to lack (Coombe, et al. 2012), we know little about teachers' perceptions about their LAL levels and their training needs, let alone have a comparison of different educational contexts. The present study investigated English language teachers' beliefs on assessment and, more specifically, perceptions of their LAL levels and training needs in the area and comparing these across the educational contexts of Germany and Greece.

114 teachers from Germany and 379 teachers from Greece completed a survey questionnaire. The data were analyzed through a series of RM ANOVAs, correlation analyses, and confirmatory factor analysis. The results indicated that teachers share the same beliefs and concepts of assessment and LAL, but their practices as expressed in their training levels and needs differed depending on their educational contexts. The poster problematizes the importance of context in assessment and offers recommendations for culturally responsive assessment literacy enhancement programmes.

Towards increased authenticity in integrated writing tasks: Construct operationalization in rating scales

Santi Lestari

Integrated tasks such as reading-to-write tasks are more commonly used in tests nowadays due to their perceived benefits, i.e. increased fairness and authenticity. Early research on integrated writing tasks concentrated on the comparability between integrated and independent tasks of scores and discourse features of test-takers' responses (e.g. Cumming et al., 2005). While this early research tended to focus on the productive skill involved, i.e. writing, a growing body of research has also investigated the role of the receptive skill(s), i.e. reading and/or listening (e.g. Plakans, 2009b, 2009a), and test-takers' cognitive processes during task completion (e.g. Chan, Wu, & Weir, 2014) to better define the construct of integrated writing tasks. Findings support that this task type elicits cognitive processes beyond what independent writing tasks do. However, research on how this distinct construct is operationalized in rating scales remains limited. It is of critical importance, however, that the construct is adequately operationalized in all aspects of the test, including in rating scales; otherwise, serious doubts on the authenticity, and even more so, the validity of the score interpretations and uses will arise.

This poster will present on-going research on how different rating scales attempt to operationalize the construct of integrated reading-to-write tasks, and how they might facilitate raters to better operationalize the construct while rating. The poster will describe the research design, including a mixed-methods approach, and initial pilot study findings. The study hopes to increase our insights into construct representation in rating scales, thereby helping ensure authenticity in integrated skills testing.

Assessing low-level writing: Identifying the need for change

Veronika Schwarz, Franz Holzknecht, Eva Konrad and Carol Spöttl

Research on assessing writing is traditionally concerned with higher proficiency levels, while studies on lower levels are sparse. However, it can be argued that the practical need for assessing low-level writing is increasing, not least due to a heightened importance of language assessment for migration. The CEFR companion volume published in 2018 addresses this need by including more descriptors at the lower levels. In this respect, the extended CEFR can be seen as a transition point in foreign language education. However, it is unclear how the new descriptors will affect writing assessment practices.

This study investigated the theory and practice of low-level writing assessment. We applied a mixed methods design consisting of three parts: a detailed analysis of 45 sample tasks at A1, A2 or comparable levels from 21 writing exams by 10 international test providers; an online survey about current practices of 42 test developers; and an analysis of the extended CEFR descriptors to investigate whether these might be helpful in addressing some of the identified challenges.

The results revealed that the literature currently offers little guidance for developing low-level writing tasks. In addition, the task analysis and survey results showed that a clear differentiation between A1 and A2 tasks is lacking, leading to large differences in what is being assessed at these proficiency levels. We argue that the extended CEFR descriptors may be useful in addressing some of these shortcomings, particularly with regards to increasing authenticity in low-level writing assessment.

Striving for authenticity in testing listening

Doris Frötscher and Nikolaus Giffinger

Authenticity is a desirable feature of input materials in listening assessment and plays a crucial role in cognitive and context validity (Elliott & Wilson, 2013; Field, 2013; Ockey & Wagner, 2018). In the context of our national examination, two kinds of input materials for assessing listening at CEFR levels B1 and B2 are used: a) authentic broadcast material, and b) spontaneous, non-scripted or semi-scripted monologues or dialogues by native speakers, recorded by item writers.

Combining the assessment professional and the sound technician perspectives, this poster provides insights into the steps we follow in our test development cycle in operationalizing authenticity in testing listening. On the assessment side, these steps comprise theoretical and practical training of item writers to find or record sound files, one key aspect being features of spoken language in the material used. On the technical side, our development cycle includes a sound file check where an objective list of quality criteria ensures that only good audio material goes forward to task production. Details on the acoustic and technical criteria analyzed and the respective acceptability boundaries are reported. Furthermore, statistics on the causes for sound files to be rejected are presented (n=322 sound files over 1 year).

We will illustrate how personalized feed-back to item writers and guidelines on how to (better) record sound files have drastically reduced the rejection rate of sound files despite stricter quality standards. Positive knock-on effects on subsequent phases of the test development cycle are discussed.

Large-scale test accommodations: From practicality to a research and validation agenda

Richard Spiby and Judith Fairbairn

As governments and institutions pass new laws to increase accessibility for test-takers, the challenge for test developers is to ensure that tests meet both accessibility and validity criteria as specified by major language testing associations (ILTA, 2007; AERA, 2014). Relatively little research has been conducted in this area of second language assessment (Taylor, 2012). Considering the centrality of test-taker characteristics to models of test validity (O'Sullivan, 2011), and consequently the disparate nature of test accommodations provided, it is unsurprising that test providers face difficulties in obtaining evidence to validate test modifications.

This poster deals with accessibility in tests with reference to both operational testing and validation research. Using practical experience gained within the organisation and using principles applied from the literature, we have pursued our public commitment to mainstreaming equality, diversity and inclusion in our approach to testing. Accessibility is addressed at all stages of the test development cycle informed by consultation with a variety of stakeholders and this has now generated a set of case studies which can be used to investigate the validity of accommodations and their impact on test-takers.

Examples and case studies are discussed in terms of how they might contribute to a model of validity in special needs testing. With the challenges associated with the generalisability of such data, these are considered in terms of the applicability of grounded theory and ethnographical analysis and the implications of these approaches for validation of test accommodations.

Conceptual and practical challenges in assessing young learners' foreign language skills

Ari Huhta and Karoliina Inha

Our poster demonstrates some of the conceptual and practical challenges in designing context-specific language assessments and other data collection procedures for young, 6–9-year-old children. The assessment tools enable us to study how children's (foreign) language proficiency develops from an early age up until the end of compulsory education.

We present a longitudinal study comparing the effects of three different starting points for learning English as a foreign language (EFL): starting in grade one (age 6-7), grade two, or grade three. We cover two stages of the study: stage one (spring 2018) and stage two (planned for spring and/or autumn 2019). We first present the design and main results of stage one, in which approximately 150 first, 300 second and 300 third graders participated. All children had studied English for the previous six months but had begun their English studies in different grades (grade one, two or three). We conclude with describing the research design of stage two.

The challenges of stage one included the need to use (for practical reasons) mostly written tasks with learners whose EFL teaching focused on developing oral skills, and whose first language reading and writing skills were still developing. Another challenge concerned deciding the focus of assessment, as only a very limited time was available for data gathering. As for stage two, our concerns relate to widening the scope of language skills assessed while maintaining links with the assessments carried out in stage one to enable tracking how the learners' language proficiency develops.

Classroom-based assessment of oral mediation: Challenges and opportunities

Olga Lankina and Yulia Pets

The presentation examines the classroom-based assessment of Oral Mediation (mediating texts and concepts by means of one language, English). Even though collaborating in a group and leading group work are extensively used in the modern classroom, it is still a problem to find an adequate way to assess mediation integrated into a conversation. The objective of this presentation is to show the advantages and challenges of using group discussions for formative and summative assessment and suggest that it can be done by giving candidates Global Achievement marks for mediation on a par with Analytical marks.

The empirical research underlying the presentation covers all test stages and features undergraduate students of two universities (ca.100 people, B1 – C1 CEFR). The instruments that are being used include classroom observation, results analysis with classical methods of statistics (ITEMAN) and the many-facet Rasch model (FACETS). The procedure of the oral test involves written or video input, preparation, group discussion (3 – 4 students) and self/peer assessment. Marks awarded by professional raters are mapped onto peer and self-assessment.

The practical relevance of the presentation is that it offers a test in the format which is commonly used in the classroom and provides criteria for assessing mediation when it is part of a group discussion. The sample FACETS specification for oral tasks makes it a useful tool for a reliable and efficient analysis. The challenges of the method, including group format and reliance on peer and self-assessment, are discussed.

Finding equal balance between standardised tests and classroom-assessment

Yevgeniya Pronoza

The poster session will talk about how our language center (at a Gulf government university) successfully transitioned towards having a fine balance between standardized and classroom assessment. It's a case where top-down (institutional policy) and bottom up (teacher expertise) initiatives have met and resulted in an equal distribution of assessment of learners' abilities through paper-and-pencil tests and alternative (classroom-based) assessment methods.

First, the poster will outline the old practices – where standardized testing prevailed in some programs, while others were dominated by classroom assessment and teacher judgement. Next, I'll have a section outlining pros and cons of each of the assessment traditions and their effect on student progression. A third section will be devoted to how this problem was overcome and what we as a language program did to ensure the smooth transition towards the balance of these two (needs analysis, revision of student results, teacher training, assessment documents preparation). The final section will present the present composition of the variety of assessment methods in different programs and outline the way for future improvement.

The language center I work at is one of the biggest in the Gulf and offers around 30 language courses of all levels ranging between low to high general English proficiency levels and a number of ESP courses. Ensuring all our language programs follow a standard assessment procedure is quite a challenging task and very interesting to share.

Reading for success: Investigating readers' cognitive processes in Austrian EFL reading tests

Klaus Siller and Andrea Kulmhofer-Bommer

The nationwide assessment of English as a foreign language (EFL) in Austria in 2013 (E8 reading test) revealed that a significant number of students cannot read sufficiently after four years of learning EFL. As previous research has shown (Siller, 2017), this lack of reading comprehension skills is closely related to the cognitive and metacognitive processes (Khalifa & Weir, 2009) at play when dealing with reading comprehension tasks. Hence, it is of utmost importance to investigate how (un)successful students deal with such reading comprehension tasks to learn more about the students' skills and the test. Such findings will fuel changes in assessment literacy and transitions in test development alike. Hence, this poster presentation shares the theoretical framework as well as the initial research design behind a research project conducted in Austria dealing with identifying students' approaches to solving reading comprehension tasks. The intended goals of this project are to learn more about the reading processes and strategies adopted by students in order to improve test developers' understanding of the particular tasks. Such findings should then inform necessary transitions a test such as the E8 reading test needs to undergo in order to improve cognitive validity. The novel perspective adopted by looking at how students arrive at incorrect solutions will shed a new light onto the test and the tasks. Participants are invited to discuss and challenge the research design as well as the intended goals and their contribution to the field of language testing.

Scientific conference: An authentic environment to assess medical students' English communication skills

Eva Braidwood and Magdalini Liontou

Simulation is a common methodological practice in medical teaching. Similarly, its counterpart, role-play is a useful tool in the language classroom. In medical English courses such tasks are used primarily to support students in developing doctor-patient communication skills. In our new English medical communication course we take simulation further; it informs the overall course design and is also used for assessment.

The final course assignment is participating in a simulated student conference. The course is compulsory for first year medical students of the University of Oulu, Finland, altogether 200 participants. To enhance subject learning and develop the relevance of English language skills, the ESP course has been integrated with clinical psychology. This provides the theme of the conference. The conference design allows for using various ways of assessment: peer-feedback, expert's feedback, feedback from the language teacher and finally, self-assessment. Students working in teams prepare a poster, present them to their peers and invite questions, which then lead into brief discussions—similar to a medical conference. This promotes insight into various topics presented and 'real-life' professional communication skills. We use technology (QR codes, digital platforms) to encourage peer-feedback, which we also find useful as part of their final assessment.

This poster will report the outcome of this novel assessment and present the course design together with students' reaction to the authenticity of the assignment. We have used mixed method design in evaluating students' feedback and assessing students' satisfaction.

Language assessment literacy in a Saudi context

Arwa Alyami

Due to changes in the educational system for schools in Saudi Arabia, Saudi language teachers are experiencing a radical shift from the traditional-form based language curriculum to a more communicative-focused curriculum. Consequently, teachers are experiencing new roles of classroom-based assessment practices, and they are expected to use new types of assessment procedures to assess and improve their students learning. To support teachers to cope with the current change, there is an urgent need to focus on their language assessment literacy. This paper is part of an ongoing research study which is set out to examine the language assessment literacy of Saudi teachers and the place of assessment in the continuous professional development programs. It focuses on some major issues in the literature of language assessment literacy which have been addressed but yet to be solved; These include, what exactly is meant by language assessment literacy? And What is the role of teachers' identity and context in defining this literacy?

14.00-15.15 **Parallel Works-in-Progress**

George Moore Auditorium

14.00-14.25

Developing an L2 speaking test corpus: Construction and analysis of a pilot corpus

Luke Harding, Dana Gablasova, Vaclav Brezina, John Pill and Jamie Dunlea

One recent transition in language assessment is the increased recognition and integration of methods from corpus linguistics within assessment research practices (Cushing, 2017). Corpora can provide important insights at the test development stage and can also provide valuable data for validation research. This work-in-progress presentation will describe a collaborative project between a team of academic researchers and an international testing organisation to construct a corpus of speaking performances drawn from a multilevel English language test. The project involves two stages: a) construction of the pilot version of the speaking corpus; b) corpus-based analysis of language use in speaking test tasks.

The corpus will include 900 exam candidates' performances on the four tasks which make up the speaking test. Candidates will be drawn from three proficiency levels – B1, B2 and C1 of the CEFR – and from three linguistic backgrounds (Spanish, Arabic and Chinese) balanced across the proficiency levels. It is estimated that the overall corpus size will be approximately 540,000 words, with the overall size of each sub-corpus at approximately 180,000 words. In the presentation we will first describe the construction of the corpus, from transcription to annotation. We will then show findings from preliminary corpus analyses of performances and specifically discuss initial explorations of the linguistic features which characterise language functions within the corpus. Finally, we will describe further plans for the development and use of the corpus, inviting discussion from the audience about next steps.

14.30-14.55

Assessing teacher discourse in a spoken English proficiency test

Odette Vassallo, Daniel Xerri and Larissa Jonk

Teacher discourse is often overlooked during pre-service teacher education, and trainees are expected to develop this competence while on the job. Feedback from the monitoring of standards in private language schools in Malta revealed that both newly qualified teachers and seasoned practitioners lacked awareness of the use of context-specific language as classroom discourse. Consequently, teachers were often unclear in their explanations, instructions, feedback, etc., and failed to offer a good language model for their learners. With the introduction of the Spoken English Proficiency Test for Teachers (SEPTT), teacher discourse became one of the criteria adopted for assessing the spoken production of pre-service teachers. This paper presents the work-in-progress of a study that analyses the spoken corpus data generated by this innovative test. It describes the challenges involved in designing the instruments used to replicate classroom tasks and routines in order to simulate teacher discourse for candidates. The paper discusses the significance of a teacher-specific register to pre-service education (Freeman et al., 2015; Van Canh & Renandya, 2017) and illustrates how SEPTT embodies a context-specific approach to language assessment as a means of resolving problems with teacher discourse. Key preliminary findings demonstrate how within a year of the test's launch there has been a marked improvement in the quality of teacher discourse among candidates and invites a discussion on how to improve the test.

15.00-15.25

Diagnostic-dynamic assessment - conceptual and practical innovation in foreign language assessment
Matthew Poehner, Ari Huhta and Dmitri Leontjev

Diagnostic and dynamic assessment have both attracted considerable interest in the L2 field in recent years but interfaces between them have yet to be explored. Diagnostic assessment, rooted in cognitivist orientations to development, emphasizes carefully defined and measured constructs for the purpose of providing detailed information regarding learner strengths and weaknesses. Dynamic assessment, in turn, is rooted in Vygotsky's Sociocultural Theory and takes a dialectical view of development according to which the provision of support (mediation) through a teaching element is an integral part of the assessment procedure as learner responsiveness is interpreted as evidence of emerging abilities.

This presentation reports a work in progress intended to design and apply assessment procedures that include features of both frameworks. At a conceptual level, the differing theoretical orientations of diagnostic and dynamic assessment suggest that their integration needs to take account of how constructs are defined, development is understood, and assessment results are interpreted. Hence one aim of the project is to investigate the theoretical basis for a unified construct encompassing both frameworks. At a practical level, the impact of a novel dynamic-diagnostic assessment procedures on learning outcomes and learners' and teachers' beliefs and practices must be understood. Toward this end, both qualitative and quantitative data are crucial, including learner unassisted and mediated performance on assessments, classroom observations, and teacher and learner interviews and questionnaires. Specific details of the planned empirical study will be shared, and we invite audience discussion concerning our conceptual and practical aims and the planned research design.

E0.01
14.00-14.25

From language class to higher education: Assessing refugees in transition
Anika Müller-Karabil and Claudia Harsch

Transition to higher education is challenging, not only for home students, but even more so for international students. Here, refugees represent a recently growing, yet still under-researched group possibly facing unanticipated challenges during this transition, in terms of assessment and beyond. This Work-in-progress gives insights into the experiences of a group of refugees entering university in Germany. The participants were accompanied during a preparatory language programme and through their first academic year. Following a longitudinal mixed-method approach, the study sheds light on their transition phase from language classes to academic lectures, from language assessment to academic assignments.

The following research questions were addressed: (1) How do participants of the language programme and their language teachers/tutors perceive their linguistic preparedness for academic requirements? (2) How do they cope in their first academic year and does language assessment during the preparatory programme (achievement tests, university entrance test) predict academic preparedness in any way?

In order to answer RQ1, we draw on data both from a questionnaire amongst participants (n=41) and from interviews with participants (n=18) and language teachers/tutors (n=10). For RQ2, preliminary insights into a longitudinal interview study with 18 students will be given and combined with language assessment results and academic grades.

The work-in-progress presents preliminary results and discusses in what ways the language programme and its assessment tools fostered and informed the transition of refugees into academia, and what sort of additional support is further needed to ease this transition.

14.30-14.55

German high school students‘ preparedness for English-medium Bachelor programs

Christine Ringwald

Due to the Bologna reform, the number of English-medium instructed study programs in Germany has risen from 65 in 2001 to currently 1,438 (Gürtler & Kronewald, 2015). This remarkable increase can be observed primarily with regard to Master’s programs; however, the trend towards English-medium instruction (EMI) is also gradually spreading to Bachelor’s programs (currently 226). The transition from German secondary school to higher education is regulated by an exam (Abitur), which encompasses one module on a foreign language. Although English language proficiency is one major factor influencing EMI success, there is little research on the relationship between high school English instruction and students’ preparedness for EMI (Dimova & Hultgreen, 2015). Taking up this angle, the study examines whether passing the English Abitur module implies students’ readiness to enter EMI courses in Germany.

The study targets upper secondary English teachers, lecturers teaching in the first semester of EMI Bachelor programs, and first-year students who entered an EMI program with the Abitur. Employing a mixed-methods approach, data is collected from participant observations, document analyses, interviews, and questionnaires.

The work-in-progress discusses preliminary results from: participant observations in five EMI Bachelor programs (October 2017), five interviews with lecturers (early 2018), and five interviews with English teachers (autumn 2018). My findings should yield insights into whether the English language competences acquired at the end of secondary schooling match the academic language requirements at the beginning of EMI programs, thus contributing towards identifying support measures that may be needed to ease this transition.

15.00-15.25

Use of feedback practices at language classes in Finnish upper secondary schools
Toni Mäkipää

Research indicates that Finnish language teachers use traditional, summative assessment methods, such as exams. This poses a threat for the Finnish education. To gain greater insight into the enhancement of formative assessment, I study how Finnish language teachers give feedback at different CEFR levels (language proficiency levels) in upper secondary school. To illuminate this question, two sub-questions are examined; RQ1: What differences exist across languages regarding feedback at CEFR levels? RQ2: How do students want to receive feedback at different CEFR levels? The key constructs applied are formative assessment, assessment literacy, corrective feedback, and sociocultural learning theory. The participants are students (N=400) studying English, Swedish, and French in two upper secondary schools, and their teachers (N=8). All the participants answer an online survey in November 2018, and I will interview 12 students and four teachers showing different answer profiles in January 2019. I will apply mixed-methods approach to data analysis. Prior to answering the survey, the students write an essay; based on this essay, their teachers and I will determine the students' CEFR levels. In the conference, I will discuss the preliminary results. Feedback in English is presumably more detailed, as students should reach a higher proficiency level in English. My study is connected to assessment literacy following changes in assessment approaches, and challenges in classroom-based language assessment. The novelty lies in the research task, as I examine feedback practices across different CEFR levels and languages. The results provide language teachers with practical knowledge of adjusting feedback to students.

H1.51
14.00-14.25

Technology-assisted scoring of short-answer items for listening comprehension: A clustering approach
Leska Schwarz and Christian Gold

In this work-in-progress, we present empirical research on how to support human raters when scoring short-answer items given by learners of German as a foreign language in response to listening comprehension prompts. Short-answer questions are a popular item type, where test-takers are able to write free-form answers. Responses are scored based on their content, ignoring language errors if they do not impede comprehensibility. Responses are typically scored manually by human raters, a process that can be very time-consuming, especially in large-scale assessments.

In this study, we examine how human raters can be supported by means of language technology. In particular, we cluster responses so that similar responses can be scored consecutively. Because spelling variation is one of the main factors responsible for the variety of answers, especially for non-native speakers in listening comprehension, these clusters are formed according to the surface similarity of responses. We compare two conditions of presenting responses to human raters: A baseline where answers are presented in random order and one where responses are presented as clusters of similar responses. To date, we have collected evidence based on three different test versions, each containing 13 prompts. Every test version was administered to approx. 200 test takers. Preliminary quantitative data analysis suggests a tendency for clustering to lead to both faster as well as more consistent scoring. Qualitative feedback from raters revealed that they perceive clustered responses as very helpful.

14.30-14.55

Instruments to evaluate the development of intercultural competence
Maria Leticia Temoltzin

The Framework of Reference for Pluralistic Approaches (FREPA) is a relatively new document that establishes a list of resources and competences to guide the development of Intercultural Competence (IC) of foreign language learners (Candelier, Camillieri-Grima, Castellotti, De Pietro, Lörincz, Meißner, & Molinié, 2012). This document is a proposal that emerges from the Common European Framework (CEFR) and it serves as the basis for this research project. This presentation shows adapted instruments used in a research project to evaluate the development of intercultural competence of twelve students of English as a foreign language. These are future English teachers involved in a program based in the Common European Framework of Reference (CEFR). In this context, the development of linguistic competence is evaluated through formal evaluation, but there are no instruments to analyze the development of students' intercultural competence. As a consequence, seven instruments from Huber-Kriegler, Lázár & Strange (2003) were adapted and followed the resources in the Framework of Reference for Pluralistic Approaches, FREPA. This is a critical analysis of how the CERF, and FREPA both aimed for a European context, have been adapted by Mexican schools in order to promote the development of bicultural and bilingual abilities, knowledge, and attitudes of foreign language learners. The presentation will include the theoretical framework that supports the research, and the methodology used. Preliminary findings will be shared with the audience.

15.00-15.25

Change in policy - change in practice?
Doreen Spiteri

Pre-service teacher education represents a valuable period in the journey of becoming a teacher during which trainees learn about good practice in teaching, learning, and assessment. Skills and pedagogical content knowledge are gradually developed, among which are those relating to classroom based assessment and its pivotal role in the teaching and learning process. Pre-service teachers learn about fundamental concepts and practices in assessing learners' language skills and abilities on a daily basis in a classroom context adopting an assessment for learning approach. This preparation has recently taken on the importance it deserves following a change in local assessment policy which sees classroom based assessment made more visible and formalised as teachers are expected to record and report instances of students' learning during the school year. This evidence of student learning is intended to be used formatively to support learners with their learning, and summatively as it contributes to the end-of-year formal summative assessment. Trainee teachers on a two-year Master's level teacher education course carry out their practicum formally over a five-week period and observe lessons on one day a week during the rest of the year. This paper investigates the extent to which the trainee teachers put into practice their evolving understanding of assessment practices, what type of assessment strategies are used, how confident they were in carrying out classroom based assessment, and any factors that stood in the way of their practising their new skills. Data is collected through an analysis of their daily written reflections, lesson observations, and face-to-face interviews.

15.15-16.00 **Coffee**

16.00-17.30 **Symposia (parallel)**

16.00-17.30 George Moore Auditorium

Language tests for citizenship purposes and low-literate learners
Lorenzo Rocca, Jascha Rüsseler, Bart Deygers, Cecilie Hamnes Carlsen
Discussants: Kellie Frost, Susy Macqueen

Worldwide, 750,000,000 adults are illiterate. More than 14% of the global population does not achieve full functional literacy, while just 7% have a university degree (UNESCO, 2017). Given these numbers, it is unsurprising to find that one third of the migrants to Europe have a suboptimal educational background or low literacy (International Organization for Migration, 2017). Increasingly within Europe, these people are required to take a language test, to determine their right to permanent residence or citizenship. As such, low educated, low literate people are part of the testing population, but as it stands, the field of language testing is ill-prepared to face the challenges that come with developing or interpreting tests for this target group.

Most research in language testing and applied linguistics is premised on higher educated learners. In a context of centralized language testing for citizenship or permanent residence this presents important theoretical and ethical challenges. If testing theories are to be generalizable, they should be applicable to all language learners, not only the higher educated subset. The dearth of research on lower educated language learners thus presents an important lacuna, and not only in terms of sampling adequacy. Indeed, if we lack the knowledge and the research on an important proportion of a high stakes test population, ethical and valid testing practices cannot be guaranteed. It is unsure how it would be possible to develop adequate tests for an under researched population.

In this multidisciplinary colloquium we aim to address these issues by discussing how high-stakes language tests for citizenship and residence affect low- educated, low-literate migrants. The empirical and theoretical contributions we propose delve into national language policy, neuropsychology, and validity theory.

Kellie Frost and Susy Macqueen will be the discussants of the symposium and will rely on the arguments presented in the talks, to address whether current approaches to test fairness and validation pay enough attention to illiterate and low-educated learners. They will discuss how one might approach validating high-stakes language tests when the target population is so diverse, and the goal so contentious. At the end of the colloquium there will be ample opportunity for interaction between the audience, the presenters, and the discussants.

Current trends in language requirements for citizenship and permanent residence

Lorenzo Rocca

Before we can address a complicated issue such as using high-stakes language tests for low-educated and low-literate migrants, we need reliable data regarding the real-world situation. Collecting and analyzing such data was the purpose of the research project that is presented in this presentation.

In order to chart, compare and track the language requirements for migrants among its member states, the Council of Europe (CoE) has conducted surveys among its member states for more than a decade (2007, 2009, 2013). The primary purpose of the most recent iteration in 2018 was to map the language requirements and knowledge of society requirements for migrants. Special attention was given to policies regarding vulnerable groups, such as illiterate/low-literate migrants, women, or unaccompanied minors.

Relying on informed respondents working within the governmental structures, this survey gathered reliable information on 40 CoE member states. The data were double-checked and were triangulated using survey data provided by 15 NGOs affiliated with the CoE.

The results of this study confirm the trends from previous surveys, and show that most member states rely on language requirements or language tests to grant permanent residence or citizenship. These requirements vary from A2 to C2, but most countries require B1. Importantly, in only very few cases, are low-educated learners exempt from language requirements.

Literacy and learning to read in adulthood affect brain structure

Jascha Rüsseler

Taking the outcomes of the first presentation as a starting point, the second talk discusses how illiteracy affects the brain, and how this impacts test-taking ability. It provides results from studies with adult functional illiterates demonstrating differences in functional neuroanatomy, functional connectivity and gross neuroanatomy in reading-related brain networks in adult functional illiterates and normal readers.

Resting-state fMRI, functional MRI (fMRI) and voxel-based morphometry (VBM) were used to evaluate group differences in gross and functional neuroanatomy in 20 normal adult readers and 20 adult functional illiterates. VBM, fMRI and resting state fMRI were performed before and after seven months of intensive literacy training.

We found a number of different brain networks, e.g. the visual network and the central executive network, which showed reliable lower connectivity in functional illiterates before training compared to non-impaired adults. We could also show that participation in the literacy training resulted in changes in these networks in functional illiterates. VBM analyses revealed decreased grey matter intensities in functional illiterates before training in several reading-related brain regions compared to normal readers (e.g. superior temporal gyrus, supramarginal gyrus, angular gyrus). During word reading, training-related changes were observed in the left visual word-form area and in the left inferior frontal cortex.

These findings show that the adult illiterate brain differs in several aspects from the brain of normal readers. This should be taken into account in the development of language ability tests for poor adult readers. The presentation concludes with suggestions for literacy assessment in very poor adult readers.

Measuring in-course language gains and language proficiency among learners with varying educational backgrounds

Bart Deygers

In Flanders, Belgium, most migrants applying for permanent residence need to demonstrate A2 proficiency in oral and written skills. To reach this goal, L2 learners are streamlined into three different tracks, depending on their educational and cognitive profile. To reach A2, the fast track for higher educated learners takes 120 hours, whereas the slow track offers 1200 hours. The assumption is that after completion of their course requirements, learners from different tracks will pass the A2 level. This study tells a different story, however.
Standardized writing tasks and the Dutch version of the Peabody Picture Vocabulary Test-III (Dunn, Dunn, & Schlichting, 2005) were administered to 1063 learners at different points in the three tracks. Oral performance data were collected from 150 learners. The tests were double rated, and the performances were transcribed and coded using 18 different measures of syntactic and lexical

complexity, accuracy, and fluency (CAF). In addition, data regarding participants' demographic background, migration history and linguistic profile were collected. The analyses were conducted using Many-facet Rasch, multiple linear regression and non-parametric tests, when required. The results show significant and large differences ($d < 1$) between the performances of participants at the end of the three educational tracks. More importantly, the performance differences translate to significant and large differences in terms of pass probability on the high- stakes A2 test. A longitudinal perspective on the data shows that from the very start, learners in the three different tracks exhibit substantial differences in terms of CAF- differences, which increase over time.

Language tests to regulate migration – A validity problem
Prof. Cecilie Hamnes Carlsen

A major transition in language assessment during the past 10-15 years is the use of language tests to control migration: More and more countries introduce language requirements for entrance to the host country, permanent residency and citizenship, and the requirements get stricter by the day (CoE-surveys, 2007, 2009, 2013, 2019). With this change in politics, vulnerable learner groups (refugees, low-literate, women, minors) have become part of the test population. Voices both in the field of assessment and migration research have spoken up against such requirements, which they argue violate democratic as well as human rights. Such critical opinions are however dismissed by opponents as value statements or political proclamations, and professionals opposing such use of language tests are being accused of political activism incompatible with scientific values of objectivity and professional conduct.

In this paper, I make a professional claim that language tests for permanent residency and citizenship are invalid by default, since the construct of language tests are language, not the willingness, ability or success in the integration process. I argue that language test developers should critically examine the use of language tests to control migration from a professional stance, hence within a theoretical frame of validity as presented by Messick and Kane. Since it is the responsibility of language test developers to ensure a valid, fair and justifiable use of our tests, striving to prevent test misuse is a matter of validity and professionalism, and not one of political opinion and personal values.

16.00-17.30 E0.01

Transitions in the use of C-tests
Meg Malone, Yiran Xu, Benjamin Kremmel, Steffen Brandt, Merve Demiralp
Discussant: Claudia Harsch

Over the past three decades, both researchers and educators have developed and used C-tests for placement and diagnostic purposes because they are quick and reliable measures of language proficiency (Dörnyei & Katona, 1992; Grotjahn, 1987). New research shows transitions in C-test development to use these measures for new purposes and across different audiences. This symposium shows that C-tests are being developed and used for new purposes, including screening and research as well as with more diverse audiences than in the past, including refugees and community (heritage) language learners. This symposium examines ways to develop, use and provide validity arguments for C-tests across a variety of settings.

The symposium first provides an introduction to C-test research and development in the past decade. Next, it examines three efforts in three countries and across two continents to develop C-tests in new languages and for diverse audiences and different purposes. These papers highlight the opportunities C-tests provide for language assessment development and research, as well as the challenges provided in such efforts. The symposium will end with a discussion on how these approaches to C-test development, use and establishment of a validity argument represent a transition in language test development and use.

Chinese C-test

Yiran Xu and Meg Malone

Despite ongoing popularity of C-tests in different languages during the past decade, including Japanese (Sasayama, 2018), Korean (Son, Kim, Cho and Davis), little attention has focused on developing a Chinese C-test (Arras & Grotjahn, 1994). Building on previous work (Norris, 2018), this study explores ways to refine the test development process.

The development process included quantitative and qualitative research with different stakeholders and relied on concurrent tests and qualitative feedback from test takers to develop final forms of the test and explore a validity argument. We discuss findings regarding: (1) 25 experts' feedback on the selection and difficulty levels of 20 authentic texts, (2) two deletion methods (word- and stroke-based) applied to the Chinese script, (3) 61 Chinese native-speakers' performance on an initial version of the C-test, including think-aloud data and their qualitative feedback, and (4) 34 Chinese learners' performance on a revised version of the C-test, plus think-aloud oral proficiency and reading test performance data. Results from expert feedback demonstrated a high degree of consistency in text difficulty ratings (*ICC* = 0.86). Their feedback was incorporated into the revisions. Native speakers' performance showed high accuracy on both word- (98.87%) and stroke-based deletion methods (97.92%), while qualitative feedback revealed more challenges in the stroke-based version. Thus, the word-based deletion method was used in the. Preliminary findings showed that the C-test is a reliable tool to differentiate learners from intermediate to advanced levels. The correlations between the C-test scores and the OPI, RPT scores are discussed with research and classroom applications.

C-tests as language placement tools for displaced people

Benjamin Kremmel, Claudia Harsch and Steffen Brandt

In light of current global political developments with increasing numbers of displaced people, assessment plays a key role in enabling migrants' and refugees' access to educational resources. Initiatives to provide displaced people with language support need resources to help with placement processes in order to assess people's language abilities prior to admitting or directing them to suitable courses. Since C-tests have been repeatedly shown to be practical, reliable and useful tools for placement purposes (e.g. Eckes & Grotjahn, 2006; Harsch & Hartig, 2015), a battery of such tests targeting all CEFR levels would provide a highly valuable formative resource for language learners and users around the world.

At the EALTA conference 2017 in Sèvres, 15 language testing experts formed a group to address this real-world assessment need with the support of EALTA. They have by now developed such a multilingual, web-delivered C-test battery. Our presentation will report on this development project involving six different languages (English, French, German, Spanish, Italian, and Finnish). The process to date will be outlined, from the challenges of negotiating design principles across different languages, to the online implementation, and the founding of a charitable NGO: LaPlace – Language Placement for Displaced People. We will present initial piloting and validation endeavours for linking this multilingual testing toolbox to CEFR levels and we will provide an outlook of the next steps in the project.

A mixed-methods analysis of an online screening test for Turkish L2 learners

Merve Demiralp

Due to recent global developments, there has been an increasing number of Turkish second language (L2) learners studying at Turkish universities. In order to facilitate the admission and enrollment of international students into these universities, the Turkish Proficiency Exam (TYS) was developed. Given the need to pass TYS to study at Turkish universities, students and educators are in urgent need of a low-cost screening test that would help to determine whether a student is ready to take TYS.

This study developed and evaluated an online Turkish C-test as a screening test for TYS. Participants who have recently taken TYS (N=80) were recruited to do the online Turkish C-test and a feedback survey. Semi-structured interviews were also conducted with a focus group of test takers (N=14) and instructors of Turkish (N=5) to examine the face validity of the test. The relation between the Turkish C-test scores and TYS proficiency bandings was investigated through correlation analysis, cross-tabulation, and ordinal logistic regression. The semi-structured interviews were analyzed using thematic analysis.

Findings reported that C-test scores were moderately to strongly correlated with each TYS section as well as TYS total score. Qualitative data suggested that test takers and instructors were sceptical about the relevance of the Turkish C-test to the spoken sections of TYS, which contradicted the quantitative findings. Overall, findings suggest that the Turkish C-test can predict success or failure in TYS and could therefore be used as a screening test.

ΕΥΡΩΠΑΪΚΟ ΧΑΡΤΟΦΥΛΑΚΙΟ ΓΛΩΣΣΩΝ
EUROPEAN LANGUAGE PORTFOLIO
PORTFOLIO EUROPEEN DES LANGUES
ICC
the international language association
Panhellenic federation of language school owners
PALSO

Sunday, June 2, 2019

09.30-11.00 **Keynote Symposium** George Moore Auditorium

Validity models in transition

Since EALTA 2016, the Executive Committee, in conjunction with the local organizing committee, have invited speakers to participate in an invited symposium on a theme particularly timely and relevant to the conference and the present issues facing the field. This year the Executive Committee has proposed an invited symposium on the theme of validity models in transition. Six speakers prominent in the field will present 5 minute position statements, followed by the discussant and open discussion amongst the participants and the wider audience.

Validity is at the core of language testing and assessment. However, validity theory is itself in a constant state of change. While there are certain core principles on which validity theorists agree, within language testing and assessment there is debate around the scope of validity (particularly the inclusion or exclusion of test use and consequences), the interface between linguistic theory and validity (e.g., the changing nature of language constructs), and the preferred models by which validation of assessments might be carried out. These different stances concerning validity represent debates about the epistemology of language assessment itself. It is therefore important to articulate views and beliefs concerning validity in order to understand the changing nature of the field, and to make plain the theoretical assumptions which underlie the work we do.

In this symposium, six speakers will share their perspectives on different elements of validity. The six speakers will look at the areas outlined below.

Michael T. Kane

Test scores represent more-or-less abstract claims about test takers, and validity arguments evaluate the plausibility of these claims.

Talia Isaacs

Has validity in the field of language testing moved beyond the cult of Messick? This contribution will also problematize a view on whether language testing research "fits" under the applied linguistics umbrella depending on whether it is about validity or reliability.

Detmar Meurers

With language data becoming available in a range of language learning contexts thanks to digitization and language learning applications, when can the analysis of "language in the wild" provide valid insights into language competence?

Barry O'Sullivan

In the same way as our thinking around validity evolves over time, models of validity must also evolve to remain pertinent.

Claudia Harsch

I will take a stance from a practitioner's perspective - how can teachers in classroom contexts develop valid assessment tasks that reflect the learning outcomes and teaching goals, as well as the target language uses?

Constant Leung

We are still working towards a clearer understanding of English for Academic Purposes, particularly in respect of writing. What price ‘validity’ in assessment?

Luke Harding will summarise the common themes across the six talks and provide further questions to which the audience will be invited to respond.

11.00-11.30 **Coffee**

11.30-12.30 **Keynote 3** George Moore Auditorium

Flexibility and Utility in Assessment and Validation
Michael Kane

The design and validation of an assessment assumes some interpretation or use of the results/scores to be generated by the assessment, and an evaluation of how well the assessment is working in any context requires an evaluation, or validation, of the claims inherent in each interpretation/use. Interpretation-specific conceptions of validity assume some particular kind of interpretation/use for the scores, and they suggest particular prescriptions for validity evidence. General conceptions of validity are designed to apply to a range of interpretations/uses and provide general guidelines for developing evidence tailored to different interpretations/uses. A general, argument-based approach to validation involves two stages, a development stage and a critical stage. During the development stage, the proposed interpretations and uses are specified as an interpretation/use argument (IUA), the assessment is developed, and a preliminary case for validity is made. During the critical phase, the most questionable assumptions in the IUA are evaluated. Both assessment development and validation are guided by the claims being made.

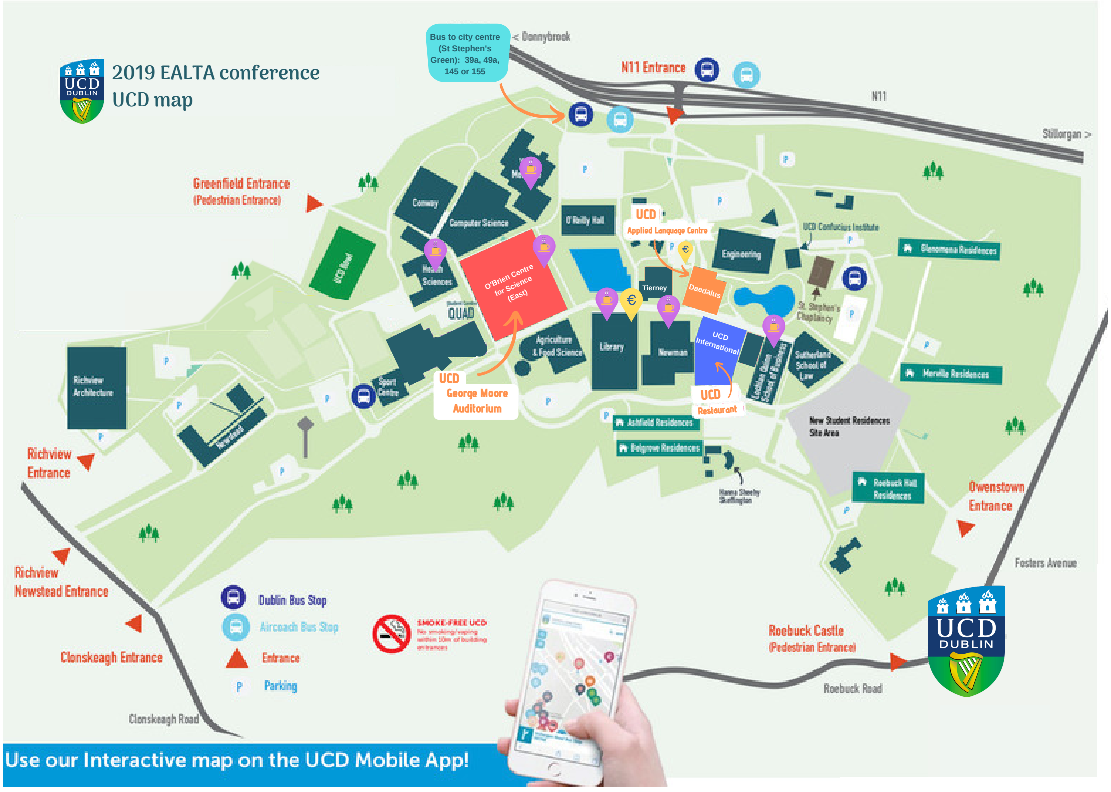
12.30-13.00 **Closing ceremony** George Moore Auditorium

Conference Venue

O'Brien Science Centre UCD

George Moore Auditorium

UCD Campus Map



UCD Applied Language Centre

List of Presenters

Arwa Alyami	arwa.alyami@ucdconnect.ie
Jayanti Banerjee	j.v.banerjee@gmail.com
Malgorzata Barras	malgorzata.barras@unifr.ch
Eva Braidwood	eva.braidwood@oulu.fi
Steffen Brandt	steffen@opencampus.sh
Tony Clark	clark.t@cambridgeenglish.org
Louise Courtney	louise.courtney@acer.org
Nivja de Jong	n.h.de.jong@hum.leidenuniv.nl
Merve Demiralp	merve.demiralp@bristol.ac.uk
Bart Deygers	bart.deygers@kuleuven.be
Lieve De Wachter	lieve.dewachter@kuleuven.be
Slobodanka Dimova	plq379@hum.ku.dk
Jamie Dunlea	Jamie.Dunlea@britishcouncil.org
Kathrin Eberharter	Kathrin.Eberharter@uibk.ac.at
Heidi Endres	endres.hi@cambridgeenglish.org
Judith Fairbairn	judith.fairbairn@britishcouncil.org
Lilian Frisen	liliann.frisen@kau.se
Kellie Frost	kmfrost@unimelb.edu.au
Doris Frötscher	doris.froetscher@bmbwf.gv.at
Nikolaus Giffinger	Nikolaus.Giffinger@bmbwf.gv.at
Elke Gilin	elke.gilin@kuleuven.be
Christian Gold	ude.christiangold@gmail.com
Tony Green	Tony.Green@beds.ac.uk
Luke Harding	l.harding@lancaster.ac.uk
Claudia Harsch	charsch@uni.bremen.de
Jordi Heeren	jordi.heeren@kuleuven.be
Raili Hilden	raili.hilden@helsinki.fi
Cecilie Hamnes Carlsen	Cecilie.Hamnes.Carlsen@hvl.no
Peter Holt	pete.holt@kcl.ac.uk
Franz Holzknecht	franz.holzknecht@uibk.ac.at
Philip Horne	philip.horne@britishcouncil.org
Chin-Ni Hsieh	chsieh@ets.org
Ari Huhta	ari.huhta@jyu.fi
Kristina Hultgren	kristina.hultgren@open.ac.uk
Talia Isaacs	talia.isaacs@ucl.ac.uk
Larissa Jonk	larissa.jonk@um.edu.mt
Michael T. Kane	mkane@ets.org
Katharina Karges	katharina.karges@unifr.ch
Gill Kendon	g.b.kenson@reading.ac.uk
Nahal Khabbazbashi	nahal.khabbazbashi@beds.ac.uk
Zoltán Kiszely	zoltan.kiszely@inyk.bme.hu
Charlotte Krämer	charlotte.kraemer@uni.lu
Benjamin Kremmel	Benjamin.Kremmel@uibk.ac.at
Andrea Kulmhofer	a.kulmhofer@bifie.at
Daniel Lam	daniel.lam@beds.ac.uk
Olga Lankina	olga_lankina@mail.ru

Steven Lattanzio	**slattanzio@lexile.com**
Dmitri Leontjev	**dmitri.leontjev@jyu.fi**
Peter Lenz	**peter.lenz@unifr.ch**
Santi Lestari	**s.lestari@lancaster.ac.uk**
Maria Leticia Temoltzin	**letytemoltzin@hotmail.com**
Constant Leung	**constant.leung@kcl.ac.uk**
Magdalini Liontou	**magdalini.liontou@oulu.fi**
Susy Macqueen	**susy.macqueen@anu.edu.au**
Toni Mäkipää	**toni.makipaa@helsinki.fi**
Meg Malone	**malonem@georgetown.edu**
Detmar Meurers	**m@sfs.uni-tuebingen.de**
Jupp Möhring	**jupp.moehring@uni-leipzig.de**
Anika Müller-Karabil	**mueller.karabil@uni-bremen.de**
Fumiyo Natatshura	**fumiyo.nakatsuhara@beds.ac.uk**
Karen Ní Chlochasaigh	**karen.nichlochasaigh@mic.ul.ie**
TJ Ó Ceallaigh	**tj.oceallaigh@mic.ul.ie**
Barry O’Sullivan	**barry.o'sullivan@britishcouncil.org**
Nathaniel Owen	**nathaniel.owen@open.ac.uk**
Yulia Pets	**j_u_lia@mail.ru**
Matthew Poehner	**mep158@psu.edu**
Yevgeniya Pronoza	**jenya@squ.edu.om**
Juhani Rautopuro	**juhani.rautopuro@jyu.fi**
Monique Reichert	**monique.reichert@uni.lu**
Christine Ringwald	**ringwald@uni-bremen.de**
Lorenzo Rocca	**lorenzo_rocca@libero.it**
Jascha Rüsseler	**jascha.ruesseler@uni-bamberg.de**
Leska Schwarz	**leska.schwarz@testdaf.de**
Veronika Schwarz	**v.schwarz@uibk.ac.at**
Prithvi Shrestha	**prithvi.shrestha@open.ac.uk**
Klaus Siller	**klaus.siller@phsalzburg.at**
John Slaght	**j.slaght@reading.ac.uk**
Paula Souza	**paulaletras@gmail.com**
Richard Spiby	**richard.spiby@britishcouncil.org**
Doreen Spiteri	**doreen.spiteri@um.edu.mt**
Carol Spöttl	**Carol.Spoettl@uibk.ac.at**
Parvaneh Tavakoli	**p.tavakoli@reading.ac.uk**
Veronika Timpe Laughlin	**vlaughlin@ets.org**
Dina Tsagari	**dina.tsagari@oslomet.no**
Aylin Unaldi	**aunaldi@boun.edu.tr**
Goedele Vandommele	**goedele.vandommele@kuleuven.be**
Alistair Van Moere	**avanmoere@lexile.com**
Odette Vassallo	**odette.vassallo@um.edu.m**
Karin Vogt	**vogt@ph-heidelberg.de**
Edit Willcox-Ficzere	**ewficzere@gmail.com**
Katrin Wisniewski	**katrin.wisniewski@uni-leipzig.de**
Daniel Xerri	**daniel.xerri@um.edu.mt**
Yiran Xu	**yx110@georgetown.edu**
Sonja Zimmermann	**sonja.zimmermann@testdaf.de**