University of BRISTOL

UKALTA
UK Association for Language Testing and Assessment

42nd ANNUAL LANGUAGE TESTING FORUM

25-27 November 2022

President’s Welcome

Dear colleagues

As the current President of the UK Association for Language Testing and Assessment I’m very pleased to welcome you to the 42nd Language Testing Forum (LTF) – our annual conference. It’s especially exciting to be able to come together face-to-face once again since our last gathering in Swansea in 2019.

After a 3-year break due to the Covid pandemic, I think we all felt a little out of practice when it came to planning or attending this year’s event! The legacy of the pandemic remains and there have been significant challenges in returning to an in person conference. Post-pandemic constraints on organisational budgets for conference attendance/sponsorship as well as rapidly rising costs of goods and services at the present time have proved major challenges. As a professional association, we are also mindful of issues of equity and access for our members and constituency, and we increasingly need to attend to environmental concerns associated with event management. So we are most grateful to Guoxing Yu and the Local Organising Committee, and their colleagues at the University of Bristol, for all their hard work to plan and deliver this year’s LTF.

We are very grateful to those organisations who sponsored LTF 2022, but given a significant downturn in sponsorship funding this year UKALTA is subsidising the lunches for delegates on both days, thus helping to keep registration fees down. UKALTA is also offering additional Student Travel Awards this year. Instead of a large-scale, sponsored gala dinner in the usual LTF tradition, we have opted for a more informal and less expensive option on the Saturday evening. For this year we strongly encouraged UK-based delegates to attend in person to try and ensure the viability of LTF as a physical event for future years. A fully hybrid event – both in-person and online - is expensive and difficult to organise, but we are providing online access for any overseas delegates who normally like to attend LTF so as to reduce travel costs and environmental impact.

The LTF organisers received a good number of quality abstracts and they have been able to include as many of these as possible by making some available for viewing online in advance of the event. We shall also have a chance to remember and pay tribute to Professor Liz Hamp-Lyons, a most respected friend and colleague, who sadly died earlier this year. So I have no doubt that this year’s LTF will be a stimulating and productive time together.

Finally, we would like to take some photographs during the weekend to use for the UKALTA website and for future promotional purposes. If you would prefer not to be included in these photographs, please let Guoxing (Conference Chair) or myself know.

Best wishes for an exciting and enjoyable LTF 2022!

Lynda Taylor

Sponsors

BRITISH
COUNCIL

CAMBRIDGE
UNIVERSITY PRESS
& ASSESSMENT

Pearson

UKALTA
UK Association for Language Testing and Assessment

Wi-fi

The University provides both ***eduroam*** and **UoB Guest** wireless services in all campus wireless locations. **UoB Guest** can also be used free of charge by members of the public from places such as our coffee shops.

How do I connect to UoB Guest?

Staff, students and visitors capable of using **edu*roam*** should do so in preference to **UoB Guest** (edu*roam* will give you a far better user experience, is much faster and gives access to internal resources).

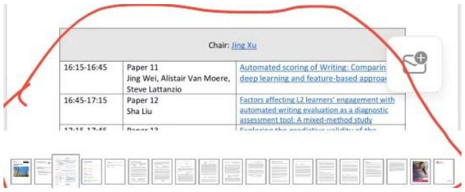
Visitors that don't have access to **edu*roam*** can easily connect to **UoB Guest**:

1. Connect to the **UoB Guest** wireless signal.
2. Your device will ask you to sign in to the Wi-Fi network.
3. You will be asked to select an authentication method – the quickest and easiest method is to use either your Google, Facebook or Twitter account. Alternatively, you can opt to receive a code via SMS text message.
4. Follow the on-screen instructions to get connected.

How to use the timetable

In computer, please open the PDF in Adobe, click the title link to view an abstract. To return to the timetable, please use Alt+Left Arrow Key at the same time.

In mobile phone, please click the title link to view an abstract. To return to the timetable, click the page thumbnails at the bottom of the screen.



Chair: Jing Xu		
16:15-16:45	Paper 11 Jing Wei, Alistair Van Moere, Steve Lattanzio	Automated scoring of Writing: Comparin deep learning and feature-based approa
16:45-17:15	Paper 12 Sha Liu	Factors affecting L2 learners' engagement with automated writing evaluation as a diagnostic assessment tool: A mixed-method study
17:15-17:45	Paper 13	Exploring the predictive validity of the

PROGRAMME

Friday 25 November 2022		
17:00-18:00	Arrival tea/coffee and Registration	4th Floor Foyer, School of Education 35 Berkeley Square Bristol, BS8 1JA
18:00-18:15	Welcome and Introduction	Guoxing Yu, University of Bristol Lynda Taylor, President of UKALTA
18:15-19:15	Cyril Weir Lecture Jennifer Rowsell University of Sheffield Supported by British Council and Cambridge University Press & Assessment	Unsettling Language Acquisition: Towards Multimodal Approaches to Language Development
19:15-21:00	Drinks Reception	Subsidised by Cambridge University Press & Assessment
Saturday 26 November 2022 Arrival tea/coffee		
Chair: Tony Clark		
09:00-09:30	Paper 1 Jamie Dunlea, Aidan Holland, Carolyn Westbrook, Johnathan Cruise	Using the CEFR to improve classroom assessment in an international language programme
09:30-10:00	Paper 2 Glyn Jones	Replicating the validation of the CEFR illustrative descriptors: preliminary findings
10:00-10:30	Paper 3 Susan Sheehan, Thuy Thai	Classroom-based assessment in Vietnam: Exploring teachers' assessment for online classes
10:30-11:00	Paper 4 Lulu Zhang	Characterising feedback cultures in UK higher education from the viewpoints of teachers

11:00-11:30	**Tea/Coffee Break** (Subsidised by Pearson)	
Chair: Alistair Van Moere		
11:30-12:00	Paper 5 Raffaella Bottini *Winner of LTF2021 Best Student Presentation Award*	The effect of topic familiarity on learners' lexical complexity in L2 spoken exams
12:00-12:30	Paper 6 Yuanyue Hao	Assessing lexical stress in L2 speech by Chinese learners of English: How many rating categories?
12:30-13:00	Paper 7 Edmund Jones, Saeid Mokaram, Jing Xu	Exploring hybrid marking for a computer-based speaking test
13:00-14:00	Poster oral presentations (3 mins x 2), see the running order on p.7 Lunch (subsidised by UKALTA)	
Writing Assessment Symposium in Memory of Liz Hamp-Lyons Chair: Tony Green		
14:00-14:15	Opening	Lynda Taylor Yan Jin
14:15-14:45	Paper 8 David Slomp (via Zoom)	Ethics of writing assessment: Implications for assessment design and use
14:45-15:15	Paper 9 Phil Smyth	The interplay of exemplars and criteria in helping students and teachers to internalise writing assessment standards in Higher Education
15:15-15:45	Paper 10 Pasha Blanda	Application of the construct of coherence to diagnostic testing in English Medium Instruction in higher education
15:45-16:15	**Tea/Coffee Break** (Subsidised by Pearson)	

Chair: Jing Xu		
16:15-16:45	Paper 11 Jing Wei, Alistair Van Moere, Steve Lattanzio	Automated scoring of writing: Comparing deep learning and feature-based approaches
16:45-17:15	Paper 12 Sha Liu	Factors affecting L2 learners' engagement with automated writing evaluation as a diagnostic assessment tool: A mixed-method study
17:15-17:45	Paper 13 Duygu Candarli	Exploring the predictive validity of the Linguaskill General Writing Test for written academic performance at a university
17:45-18:15	Paper 14 Juan Zhang	Using alignment-based indicators in reading-to-write tasks: Source use as a predictor of score
19:30-22:30	Dinner	Cosmo Bristol
Sunday 27 November 2022 Arrival tea/coffee		
Multimodal Constructs Chair: Luke Harding		
09:00-09:30	Paper 15 David Booth	Efficiency and fairness: Implementing multi-modal constructs in the testing of writing
09:30-10:00	Paper 16 Ardeshir Geranpayeh, Alina Von Davier	The Role of computational psychometrics in facilitating the assessment of multimodal constructs in language assessment
10:00-10:30	Paper 17 Guoxing Yu	Multimodal constructs of language assessment: From linguistic-focused multimodality to meaning-making multimodal orchestration in integrated language assessment tasks
10:30-11:30	Q & A on the recorded paper presentations See pp.5-6 for the list of recorded presentations Chair: Tineke Brunfaut	
11:30-12:15	**Tea/Coffee Break** Group Photo (Brandon Hill, weather permitting)	

Writing Assessment Symposium in Memory of Liz Hamp-Lyons (Part 2) Chair: Jamie Dunlea		
12:15-12:45	Paper 18 Aynur Ismayilli Karakoc	What do raters attend to and how are they guided by latent criteria when evaluating integrated reading-writing essays? Evidence from think-aloud protocols
12:45-13:15	Paper 19 Santi Lestari	Scoring integrated reading-into-writing performance: Approaches to construct operationalisation and their potential impact
13:15-13:45	Paper 20 Tony Green	Evolution of L2 writing assessment + Wrap-up: Writing Assessment Symposium
13:45-14:00	**Awards, Farewell and LTF2023** Best Student Presentation Award: Sponsored by Duolingo Best Poster Award: Sponsored by TextInspector Student Travel Awards: Sponsored by UKALTA	
14:00-15:30	Lunch (Subsidised by UKALTA) 4th Floor Foyer, 35 Berkeley Square	

Recorded Paper presentations

(See pages 24-34 for the abstracts)

Recording 01

Teachers' language assessment literacy in the digital age: the construct and the affecting factors

Jing Zhang	University of Bristol

Recording 02

What can eye-tracking tell us about teachers' ratings of EFL students' texts?

Ebtesam Abdulhaleem	King Saud University
Suhad Sonbul	Umm Al-Qura University
Dina El-Dakhs	Prince Sultan University

Recording 03

Test-taking mediums as a facet of performance?

Gergely Dávid Eötvös	University Budapest

---------- multimodal constructs in language assessment --------

Recording 04

Young language learners' cognitive processes of digitalised picture-based causal explanation speaking tasks: An eye-tracking study

Wenjun Ding	University of Bristol

Recording 05

Assessing the language aptitude of YLs: multimodality and digital tools

Jasenka Čengić	University of Zagreb

Recording 06

Are plan/map-labelling tasks in listening assessments simply testing listening?

Sharon Wong	University of Bristol/Chinese University of Hong Kong

Recording 07

Exploring multimodality in L2 assessment through the lenses of eye-tracking: A systematic review

Xin Hu	Nanyang Technological University
Vahid Aryadoust	Nanyang Technological University

Recording 08

Examining the visual semiotics in a computerized multimodal listening assessment: The role of gaze and mouse-clicking

Yue Qiu	Nanyang Technological University
Vahid Aryadoust	Nanyang Technological University

Recording 09

Beyond the four skills – Using modes to assess integrated tasks and academic literacy

Chris R Smith	University of Edinburgh

Recording 10

From paper handouts to interactive websites: technology for students' self-assessment in the language learning classroom

Geraldine Bengsch	University of York

Recording 11

Exploring the role of listening comprehension in listening-to-summarize tasks: Preliminary Findings

Mikako Nishikawa	Nagasaki University
Yukio Horiguchi	Kansai University
Guoxing Yu	University of Bristol

Posters

(See pages 35-36 for the abstracts)

A Critical Examination in the General English Proficiency Test (GEPT) Intermediate Writing Test

Li-Chung, Chang University College London

Imagined Competencies: A Case Study of Washback in regard to Student Attainment, Teaching Practice, and Language Ideology from Standardised Testing at Public Universities in Southern Mexico

Alexander Black University College London

ABSTRACTS

Cyril Weir Lecture on Friday 25 November 2022

Unsettling Language Acquisition: Towards Multimodal Approaches to Language Development

Jennifer Rowsell, University of Sheffield

Despite widespread adoption of multimodal pedagogies during the global pandemic, language education remains wedded to words and linguistic systems to frame teaching methods and forms of language assessment. Digital writing draws on different modes and resources, embeds multiple media, takes various forms that are published instantly and shared widely. Digital and multimodal forms of communication demand not only new types and varieties of language skills, but also new ways of thinking about and approaching text understandings and composition. For this Cyril Weir invited talk, I will unsettle what we know about language acquisition to reimagine what it might look like within multimodal paradigms. Drawing on research studies that I have conducted over the years, I will profile ways of rethinking language acquisition through modal learning and teaching. It is time to put language education in active conversation with digital, multimodal, and participatory approaches to language development to move outside of the sarcophagus of tighter linguistic framings of what language development is and should be.

Jennifer Rowsell is Professor of Digital Literacy at the University of Sheffield. Her research interests include multimodal, makerspace, and arts-based research with young people; digital inequalities research; and, applying posthumanist and affect approaches to literacy research. She has worked and conducted research in Australia, Canada, the United Kingdom, and the United States. Her most recent co-authored books are *Living Literacies: Rethinking literacy research and practice through the everyday* (MIT Press) with Dr. Kate Pahl (Manchester Metropolitan University); *Unsettling Literacies: Directions for Literacy Research in Precarious Times* with Claire Lee, Chris Bailey, and Cathy Burnett; and *Maker Literacies and Maker Identities in the Digital Age: Learning and playing through modes and media* (Routledge) with Cheryl McLean (Rutgers University). She is Lead Editor of *Reading Research Quarterly*. She is Co-Editor of the *Routledge Expanding Literacies in Education* book series with Kate Pahl (Manchester Metropolitan University) and Carmen Medina (Indiana University) and she is a Co-Editor of *Digital Culture and Education*.

Abstracts of In-Person Paper Presentations

Saturday 26 November 2022

09:00-09:30	**Paper 1** **Jamie Dunlea, Aidan Holland, Carolyn Westbrook, Johnathan Cruise** *British Council*	**Using the CEFR to improve classroom assessment in an international language programme**

In 2022, a group of international organizations published Aligning Language Education with the CEFR: a Handbook. The Handbook emphasises the need for evidence-based alignment beyond the focus on linking tests to the CEFR which has dominated much of the CEFR-related alignment literature. This presentation reports on a researcher-teacher collaboration to implement ideas in the Handbook to improve a large-sale EFL language program offered for secondary school-age students. The programme is structured around sets of lessons targeting a CEFR level. Each set includes thematically linked tasks and activities culminating in a project. Approaches to alignment often used with assessment were adapted to build an evidence-based mapping of the materials used in the program to relevant parts of the CEFR Illustrative Scales, with the mapped descriptors then forming the basis for developing classroom assessment instruments. In stage one, a group of classroom teachers were trained to map CEFR descriptors to activities in the lessons. In stage two, the research team used the mapping to produce assessment instruments for use in class by teachers, with teachers using a checklist approach to observe and evaluate student performance on the projects. In stage three, these assessment tools were piloted in a small number of teaching centres with lower-secondary classes at B1 level. An important focus of the researcher-teacher project team has been to achieve an appropriate balance between detail and feasibility for teachers working in a busy language-school context with limited time for preparation and marking of assessments outside of class time. The presentation will describe stages one and two, with a particular focus on the process of ensuring a principled and documented approach to mapping CEFR descriptors to lessons, and will provide results of piloting in stage three, with the implications for assessment in the program going forward.

09:30-10:00	**Paper 2** **Glyn Jones** *Lancaster University*	**Replicating the validation of the CEFR illustrative descriptors: preliminary findings**

Given the significance and reach of the CEFR, the original research undertaken by North (1996, 2000) to calibrate many of the CEFR’s illustrative descriptors requires wide replication to (a) check the reliability and validity of initial calibrations, and (b) address limitations observed in the original study. In this paper, I report initial findings from a large-scale, partial replication of North's original research. 471 teachers were recruited from over 40 countries to participate in an online rating exercise modelled on North's (1996, 2000) methodology. Each teacher rated two of their learners using a checklist of 40 descriptors randomly allocated for each teacher from a pool of 368. A subset of teachers also submitted samples of written work produced by the same two learners they envisaged. These samples (n=480) were distributed to other participating teachers, who rated them against descriptors. This allowed for an overlapping design whereby a substantial subset of learners were assessed by more than one teacher. Teachers' ratings were analysed with many-facet Rasch measurement (FACETS) to obtain measures of learner ability, teacher-rater severity, and descriptor difficulty. Following data cleaning, two series of analyses were conducted: one with difficulty measures unconstrained, and one with certain descriptors anchored to North's original measures. Findings revealed that, when analysed independently, the descriptors were placed on a plausible linear scale. When anchored to North's original values, the majority of descriptors were placed at the same level as in the published CEFR, or at an adjacent level. The correlation between descriptor measures and North's original values was r=.80. However, a number of CEFR descriptors for written production not calibrated by North were found to be significantly more difficult in the replication analysis than is implied by their position in the CEFR. Implications are drawn for the importance of replication in language testing research.

10:00-10:30	**Paper 3**	**Classroom-based assessment in Vietnam:**
	Susan Sheehan	**Exploring teachers' assessment for online classes**
	University of Huddersfield	

Research suggests that teachers' and students' engagement in assessment activities contributes to students' learning development by influencing both what and how students learn. While research exists on how teachers engage in formative assessment in traditional classroom settings, little work has examined the nature of assessment in online classrooms (Krishnan, Black and Olson, 2021). This paper reports on a project which explored classroom-based English language assessment practices, including digital assessments, in Vietnam. The project aimed to promote fairness by improving access to high quality assessments as effective assessment can support and promote learning (Berry et. al., 2019) and as Crusan et al. (2016) point out it is the students who lose out if assessment practices are poor.

A mixed-method approach to data collection was adopted to understand the current assessment landscape. The methods include questionnaire (N = 2569) and teaching observations with follow-up interviews (N =7). The participants came from all stages of education and from both state and private provision. The questionnaire included questions on teachers' experiences of online assessment. Seven lessons were observed using an observation schedule developed by Sheehan and Munro (2017, 2019). To develop an understanding of the thinking behind the assessment activities observed, the teachers were interviewed. The observed lessons were in online classrooms and the impact of the pivot to online teaching and assessing was a significant issue for the participants. The preliminary results of the project revealed what were behind teachers' choices of assessment tasks in their classroom, the use of assessment feedback and the challenges faced by the teachers in their practices in online environment. The project has extended an understanding of the current assessment landscape in Vietnam. Discussions on how the data has informed an online toolkit which will provide assessment and testing resources and recommendation for best practices will be presented.

11:00-11:30	**Paper 4** **Lulu Zhang** *University of Southampton*	**Characterising feedback cultures in UK higher education from the viewpoints of teachers**

Feedback, as a source of information about students' strengths and weaknesses, is a key element in writing improvement (Bailey & Garner, 2010; Chandler, 2003). With the growing diversity in higher education, sociocultural factors have been found to have an impact on stakeholders' perceptions of and uses for feedback (Winstone & Carless, 2020). In order to optimize the benefits of feedback, it is vital to learn more about feedback cultures from various research settings. This study explores feedback cultures in postgraduate-level education and linguistic programmes from teachers' perspectives in the UK. The investigations are unfolded from three interactive levels: micro, meso and macro. The micro-level examines how teachers perceive and construct feedback; the meso-level analyses how teachers involve students and manage relations with students in the feedback process; the macro-level investigates the construction of the feedback environment. The overarching goal is to get a more nuanced understanding of feedback cultures and any potential sociocultural variations that might explain why teachers in the UK provide feedback in the way they do. This study employs a sequential mixed-methods design, featuring multiple iterations of data collection and analysis with multiple types of participants. The questionnaire (n=81) and the follow-up interviews (n=8) were designed to gain a more holistic perspective of teachers' views on feedback and their feedback practices and an in-depth interpretation of feedback cultures. The quantitative data and the interviews demonstrated how the UK's feedback cultures were influenced by teacher demographics, teacher-student relations, and the emergence of globalised education. This presentation explored feedback cultures by discussing why feedback rarely achieves what it is meant to, written corrective feedback practices, the timing of feedback, and the boost of uptake of feedback through developing authority, partnership, and responsibility-sharing. Results of this study will help enhance better interactions and mutual understanding of feedback amongst teachers and students.

11:00-11:30	**Tea/Coffee Break** **(Subsidised by Pearson)**

11:30-12:00	**Paper 5** **Raffaella Bottini** Winner of LTF2021 Best Student Presentation Award *Lancaster University*	**The effect of topic familiarity on learners' lexical complexity in L2 spoken exams**

Lexical complexity is a key measure of vocabulary knowledge with applications to research in a range of disciplines, including language acquisition and language testing (Kyle, 2019). Research has developed a range of indices and automatic tools to analyse lexical scores in learner language, and corpus linguistics has been the main method in this area. However, few studies have investigated lexical complexity in L2 speech since few spoken learner corpora are available (Gablasova & Bottini, 2022). Also, the existing research has mainly focused on the effect of learners' L2 proficiency, while task effects have not been fully investigated yet. This study aims at expanding prior research on lexical complexity in L2 speech, exploring the effect of topic familiarity. It uses the 4.2-million-word Trinity Lancaster Corpus (Gablasova et al., 2019) based on the Graded Examination in Spoken English (GESE), which is administered by Trinity College London, a large international examination board. This study uses the new Lex Complexity Tool (Bottini, 2022) and adopts a mixed-method approach. The results show a significant effect of topic familiarity (η2 ≤ .35) on lexical density, diversity, and sophistication. These findings have theoretical, methodological, and practical implications for learner corpus research, language assessment, and language teaching.   Bottini, R. (2022). Lexical complexity in L2 English speech: Evidence from the Trinity Lancaster Corpus. Unpublished PhD thesis, Lancaster University. Gablasova, D., & Bottini, R. (2022). Spoken learner corpora to inform teaching. In R. R. Jablonkai & E. Csomay (Eds.), Routledge handbook of corpora in English language teaching and learning. Routledge. Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, description and application. International Journal of Learner Corpus Research, 5(2), 126–158.   Kyle, K. (2019). Measuring lexical richness. In S. Webb (Ed.), The Routledge Handbook of Vocabulary Studies (pp. 454–476). Routledge.

12:00-12:30	**Paper 6** **Yuanyue Hao** *University of Oxford*	**Assessing lexical stress in L2 speech by Chinese learners of English: How many rating categories?**

Many previous studies have attempted to investigate the effects of different numbers of rating categories on raters' performance, but the numbers of categories in most studies are determined a priori. This study proposes a data-driven approach to determine the number of rating categories by combining comparative judgement (CJ) and latent profile analysis (LPA), with a particular focus on assessing lexical stress in second language (L2) English speech. Speech samples elicited from 56 Chinese adult learners of English in a read-aloud task were subject to CJ by four experienced raters. In CJ, a dyad of speech samples was presented to the raters, who were required to individually choose the sample with better performance among the two. Such pairwise comparison of all speech samples generated a true score for each sample based on Rasch model. In this study, three separate sessions of CJ were conducted, each assessing 1) correct placement of lexical stress, 2) distinct contrast between stressed and unstressed syllables, and 3) vowel reduction in unstressed syllables. The inter-rater reliability was .91, .88, and .93 on the three dimensions, and Infit values of all examinees fell below 1.5. Subsequently, the true scores of the 56 examinees on three dimensions were used as the input variables in the LPA, with the aim to identify data-driven classes of examinees. Different LPA models were run to fit the data. The model with the best model fit revealed a 4-class categorisation of examinees in terms of their lexical stress performance. This study is part of a larger project that aims to assess L2 prosody of Chinese learners of English, but the preliminary results suggest a promising data-driven approach to empirically determining the number of rating categories in a rating scale.

12:30-13:00	**Paper 7** **Edmund Jones,** **Saeid Mokaram, Jing Xu** *Cambridge University Press & Assessment*	**Exploring hybrid marking for a computer-based speaking test**

Hybrid marking means combining an automated scoring system and human examiners in assessing complex, constructed responses in speaking or writing tests. We have developed a hybrid marking system for a computer-based English-speaking test. Firstly, all responses are passed to the automated scoring system, which calculates a score and other numerical parameters such as a measure of how likely it is that the candidate is speaking English. Secondly, the hybrid marking system takes these quantities and judges whether the score is likely to be accurate enough; if it is not, the response is passed to a human examiner instead. The first version of our hybrid marking system makes its decisions based on four parameters and compares these to numerical thresholds. The thresholds were chosen based on a large set of past responses that had been marked by both human examiners and the automated scoring system. Recently we have been developing a more advanced version that uses machine learning and can take a larger set of input parameters.

To help test managers understand hybrid marking, we have introduced a metric called "assessment quality measure" (AQM). Each response is assigned a value of AQM and this summarizes the output of the machine learning model. This enables the test managers to judge the approximate score accuracy based on the proportion of responses passed to human examiners. We compared the first and second versions of the hybrid marking system using the same training and validation datasets (n=2624 and 3112) and found that they performed similarly well. However, the new version is more flexible and produces more stable outputs. We will also discuss other types of hybrid marking that have been described in the literature or implemented by other testing organizations.

13:00-14:00	**Poster oral presentations (3 mins x 2), see the running order on p.7** **Lunch (subsidised by UKALTA)**	
14:00-14:15	**Writing Assessment Symposium (part 1)** **Opening** **Lynda Taylor** **Yan Jin**	
14:15-14:45	**Paper 8** **David Slomp (via Zoom)** *Assessing Writing journal/* *University of Lethbridge*	**Ethics of Writing Assessment: Implications for Assessment Design and Use**

An important feature of Liz Hamp Lyon's work in the field of writing assessment was her concern for the consequences of assessment for students. In this session we reflect on the importance of this stance and its implications. Looking forward, we consider what a theory of ethics for the field of writing assessment could entail, and we examine the implications of that theory for assessment design and use.

14:45-15:15	**Paper 9** **Phil Smyth** *University of Reading*	**The interplay of exemplars and criteria in helping students and teachers to internalise writing assessment standards in Higher Education**

It is commonplace in many higher education settings for students to have access to the criteria that they will be assessed on, and, in many cases, exemplars showing examples of standards across the range of possible grades. However, there is a tension inherent between the proclaimed transparency of marking criteria, and the potential for criteria to 'emerge' from a discussion of exemplars that may or may not be reflected in the marking criteria. This is often because different paradigms of assessment are invoked which have different conceptual bases for making decisions about the quality of student work. Drawing on data collected from twelve EAP teachers in Hong Kong, this presentation explores the interplay between exemplars and criteria within the context of EAP classrooms and EAP assessment. The data reveal that teachers find it challenging to balance their conceptualisations of assessment and criteria with their preferred practices in using exemplars. There was also some evidence that teacher beliefs, and teachers views of their students' beliefs had an impact on exemplar and assessment criteria practice. Teachers who can shift between different conceptualisations of criteria for different purposes are likely to be better placed to both mark consistently and to encourage more student involvement in their learning.

The presentation concludes by drawing implications for language testing research into moderation and standardization sessions where criteria and exemplars are discussed.

15:15-15:45	**Paper 10** **Pasha Blanda** *University College London*	**Application of the Construct of Coherence to Diagnostic Testing in English Medium Instruction in Higher Education**

This paper concerns the expansion of constructs of language assessment, and applies the construct of coherence to the design of an Integrated language assessment task facilitated by technology. This paper is an exploratory, proof of concept study aimed at considering how a test of coherence of student writing might be operationalised to satisfy needs of practicality and authenticity for English medium instruction. Three gaps in recent research on diagnostic testing are identified: communicative authenticity of tasks based on register, genre and discipline; the effect of source text cohesion in integrated task designs; effects of cultural and social differences between students. To investigate effects of genre, register and discipline, a novel operationalisation of the construct of coherence for an integrated reading-writing task is used to analyse L2 writing (n = 396) in the in the British Academic Written English corpus (Alsop & Nesi, 2009). To investigate how text cohesion might be applied as an intervention for L2 students, a corpus of academic texts (n = 41) were summarised using an automated method based on 3 machine learning algorithms. Summaries were compared using cohesion based ease of reading measures reported in previous research in cognitive psycholinguistics (Crossley, Greenfield, & McNamara, 2008). To investigate social and cultural differences between students, an integrated reading writing task was piloted with four participants, followed by a questionnaire relating to language use and proficiency. Results suggest that the use of formulaic language may be affected by source use, suggesting further research into the construct of coherence may be warranted. Application of Latent Semantic Analysis to text summarisation showed statistically significant results in sentence level cohesion (p = <.001; d = 0.78). Piloting questionnaires suggest differences between participants in digital media consumption and informal conversation in L2.

15:45-16:15	**Tea/Coffee Break** **(Subsidised by Pearson)**

16:15-16:45	**Paper 11** **Jing Wei,** **Alistair Van Moere,** **Steve Lattanzio** *MetaMetrics*	**Automated scoring of Writing: Comparing deep learning and feature-based approaches**

There are two main modeling approaches in Automated Essay Scoring (AES): feature-based and deep learning. Feature-based approaches analyze characteristics of the writing which are combined and weighted in statistical models to predict human ratings. In contrast, deep learning approaches find patterns in the writing data, but the specific characteristics or variables in those patterns cannot be explicitly uncovered (they are too "deep" in complex layers of models). The AES literature treats these two approaches as distinct dichotomous paradigms (Kumar & Boulanger, 2020), but there has been little research comparing them on the same dataset. This paper analyzes 304 EFL learners' written responses to an open-ended picture description task. Student essays were double-rated on "language use" and "task completion" traits using a 0-4 scale, with the mean of the two ratings serving as the criterion score for model training. The dataset was split into training (80%) and test (20%) sets. Two models were applied, (i) a ridge regression with regularization and (ii) a deep-learning transformer-based model using Bidirectional Encoder Representations from Transformers (BERT) text embeddings. For ridge regression, a total of 189 features were initially entered into the model and through an interactive process the final model consisted of 14 features. The results show that ridge regression (r=0.85 to 0.89) outperformed deep learning (r=0.58 to 0.66) on predictive accuracy. Further statistical comparisons were made using root mean square error, percent agreement and quadratic weighted kappa. Discussion will be made on descriptive accuracy and relevance to the language testing audience. Despite low performance on this data, there is potential for deep learning to exceed feature engineering in different contexts or with different data. Implications will be drawn on how a combination of feature-engineering and deep learning approaches can provide insights for the educational audience.

16:45-17:15	**Paper 12** **Sha Liu** *University of Bristol*	**Factors affecting L2 learners' engagement with automated writing evaluation as a diagnostic assessment tool: A mixed-method study**

Despite a growing consensus on the diagnostic potential of automated writing evaluation (AWE) in L2 writing instruction, little research has been conducted on how L2 learners engage with AWE as a source of diagnostic feedback. However, learner engagement is the key to unlock the possible benefits of AWE feedback and helps establish the link between feedback provision with learning outcomes. This study investigated how L2 learners engage with AWE feedback in the process of essay revision and explored the underlying factors that may have affected such engagement. Theoretically, L2 learner engagement with AWE feedback was conceptualised as attention allocation, cognitive effort expenditure, and revision responses. Methodologically, this study used eye-tracking, in combination with stimulated recalls, questionnaires, and reflective journals, to obtain in-depth understanding of the three aspects of feedback engagement. Twenty-four Chinese EFL learners revised their writing through Write & Improve with Cambridge, a new AWE system that generates diagnostic feedback with three different levels of explicitness. Data from multiple perspectives were collected and examined, including participants' a) eye movements, b) stimulated recall interviews, c) revisions to the AWE feedback, d) perception of such feedback, e) English language proficiency, and f) reflective journals. The results suggest feedback explicitness as a determining factor affecting learners' engagement with AWE feedback and point to the need for timely, supplemental teacher or peer scaffolding in addition to the indirect AWE feedback. The results also suggest that AWE tools need to be constantly updated to improve their feedback accuracy, as error-prone feedback may cause participants to make inaccurate amendments to their writing. Teachers, on the other hand, should help learners confirm the accuracy of AWE feedback. In addition to feedback explicitness and accuracy, participants' perception of AWE feedback, their English language proficiency, and their previous experience with AWE have been found to affect how they engaged with such feedback.

17:15-17:45	**Paper 13** **Duygu Candarli** *University of Dundee*	**Exploring the predictive validity of the Linguaskill General Writing Test for written academic performance at a university**

This paper seeks to establish the extent of the predictive validity and extrapolation inference of the writing component of the Linguaskill General test, a relatively new online modular language proficiency test. García Laborda and Fernández Álvarez (2021, p. 14) reviewed modular language tests, including the Linguaskill Test and noted that "what is still missing in many of them [tests] is a real and sound corpus of research, especially in aspects such as external validity, generalization, extrapolation and decision". By drawing on an argument-based approach to validity, this study responds to the need for research in this area by contributing to the validation argument of the writing component of the Linguaskill General test for written academic performance at a UK university. This study examined the relationship between postgraduate students' Linguaskill General Writing test scores and their written academic assignment grades over time at a UK university and the relationship of the linguistics features (lexical sophistication, cohesion and syntactic sophistication) in the Linguaskill General Writing test responses to those observed in written academic assignments at university, by using corpus linguistics techniques. The findings provide empirical evidence for the predictive validity of the Linguaskill General Writing test scores for success in written academic assignments to some extent and extrapolation inference for the written language performance in academic assignments at a UK university. Keywords: language testing; predictive validity; extrapolation.

17:45-18:15 **Paper 14** **Using alignment-based indicators in reading-to-write tasks: Source use as a predictor of score**

Juan Zhang
Zhejiang University/
University of Bristol

Developing and scoring reading-to-write (RTW) tasks are challenging because item writers and raters should attend to features of source use. The concept of alignment, initially developed in the field of psychology, and later applied in L2 research, may provide insights into addressing the conundrum of integrated writing development and scoring. This research investigated how EFL students across proficiency levels align with the source text differently in RTW tasks, how alignment-based indicators correlate with those mirroring writing quality, and to what extent alignment features of source use could explain score variance.

127 Chinese EFL college students participated in a proficiency test which includes a RTW task. They were divided into three proficiency groups according to their total scores on the test except scores on the RTW task. All scripts of the RTW task were scored by two human raters using a holistic scale, and were coded by two testing professionals for chunk-level and sentence-level alignment indicators (i.e., trigram, full- & partially-aligned sentences). Lexical and syntactic complexity were measured using Coh-Metrix and L2SCA respectively. Accuracy was represented by the proportion of error-free T-units. Kruskal-Wallis test, correlation and multiple regression analyses were performed to address research questions. Results showed that low proficiency group used significantly more trigrams and full-aligned sentences than its high proficiency counterpart. Those two indicators were found to correlate positively with the number of complex nominals per T-unit. Additionally, alignment-based indicators could account for 34% of the score variance. Together with indicators of discourse features, 60.3% of the score variance could be explained. The findings suggested that it is better for item writers to avoid giving topic sentences in source materials if RTW tasks involve summary writing. Also, raters should pay special attention to gist-laden sentences in source materials which tended to be slightly modified by low-level examinees in compositions.

19:30-22:30 **Dinner**
Cosmo Bristol

Sunday 27 November 2022

Multimodal Constructs

09:00-09:30	**Paper 15** **David Booth** *Pearson*	**Efficiency and Fairness: Implementing multi-modal constructs in the testing of writing**

Innovative test design has, to a large extent, stalled, with test publishers unable or unwilling to develop and deploy state of the art technology in high stakes assessment. This presentation examines a test of English proficiency which endeavours to meet the challenges inherent in developing multi-modal test items and scoring methods to achieve the balance between assessment efficiency and language domain coverage. High stakes testing can be a stressful experience for test takers. If enough information is collected to assign test takers' accurate test scores, then extended testing times do not add to scoring information. This presentation will report on a meta-review of the statistical analyses and content reviews conducted over time by the researchers from both internal and external to the testing organisation to monitor how efficient items are in terms of how much measurement information they provide.

The session focusses on how the use of integrated skills items in a fully computer-based test model can increase the efficiency of testing which reduces stress and anxiety for test takers enabling them to perform to the best of their ability. We explore different scoring methods and challenge current paradigms which purport to balance the scoring of language skills through separate skills papers and score weighting. We will focus particularly on writing as a skill and how multi-modal constructs can be realised in conjunction with other skills and overall language proficiency.

09:30-10:00	**Paper 16** **Ardeshir Geranpayeh[1],** **Alina Von Davier[2]** *1.Cambridge Computational Psychometrics Ltd,* *2. Duolingo*	**The Role of Computational Psychometrics in facilitating the Assessment of Multimodal Constructs in Language Assessment**

In this paper we argue that computational psychometrics is an interdisciplinary field which blends psychometric principles of assessment with machine learning approaches to assessment and learning. The similarity between the two fields resides in that in psychometrics one starts with theory driven construct definition of assessment and goes on to provide evidence (data) for that theory: Evidence Centred Design (ECD) (Mislevy, et al., 2003 and Mislevy 2018). It is very much a top-down approach. Machine learning, on the other hand, starts with the observed assessment data (candidate's performance) and tries to infer the underlying factors influencing the construct being measured; a bottom-up approach. Where the two disciplines meet is data (observation), which is somewhere in the middle. Both disciplines use data mining techniques to analyse their given data, hence both rely heavily on data science. There has been an increasing trend of using multimodal integrated tasks in language assessment, facilitated by technology. The traditional psychometric methods are ill equipped to handle this new challenge of measuring such multiple constructs with complex data. Computational psychometrics offers to take advantage of the new technology to provide a new perspective to accommodate the complexities of the multiple constructs being measured (observed). We will demonstrate how within the framework of computational psychometrics the traditional theoretical psychometric concepts blend with data-driven machine learning tools to provide a better understanding of students' learning and better assessment of the multi-dimensional constructs. We bring practical examples from a high stake's language test where the application of computational psychometrics accelerates the scalability, security and accessibility of the assessments by making the test development process more efficient, blending automatic item generation with human design and review. The new assessment results in rich data that reflects multi-construct nature of the new assessment and incorporates dependencies over complex tasks and item types.

10:00-10:30	**Paper 17** **Guoxing Yu** *University of Bristol*	**Multimodal constructs of language assessment: From linguistic-focused multimodality to meaning-making multimodal orchestration in integrated language assessment tasks**

Linguistic-modes are part and parcel of visual, auditory and spatial patterns of meaning-making in our increasingly multimodal communication. From a socio-semiotic approach to multimodality, speech and writing are "partial" means of communication. The nature of partiality of speech and writing and the complementarity between and within linguistic- and non-linguistic modes challenge the long-held assumptions of the sufficiency of linguistic modes for all communicational needs, and the common practice of high-stakes language tests that focus entirely on linguistic input and output. Facilitated by technology in task design, delivery, and completion, multimodal integrated tasks are becoming popular in language assessment. Integrated multimodal language assessment tasks are traditionally defined and operationalised with reference to the requirement of different linguistic skills in task completion. With the use of technology, we see an expansion of constructs of language assessment, with multimodal input (e.g., non-verbal materials such as cartoons, maps and graphs, and videos in listening and speaking tasks) and multimodal digital composing/ensembles/orchestration (e.g., by creating a video, drawing a picture, a map, or an outline to summarize what test-takers have read/listened) in task completion. To what extent is multimodality of communication being operationalised in high-stakes, traditionally language-focused assessment tasks, in comparison to the use of multimodality in low-stakes assessments? What roles does multimodality (in task input and output) play in completion of language assessment tasks? Is multimodality a supplementary and supportive component of, or an integral part of the construct of language assessment? What are the implications of the different roles of multimodality (supplementary and supportive vs. integral) for the focus of assessment criteria (i.e., weak vs. strong version of multimodal constructs of language assessment). In this presentation, I will try to address these fundamental questions.

10:30-11:30	**Q&A on recorded presentations**
11:30-12:15	**Tea/Coffee Break** **Group Photo (Brandon Hill, weather permitting)**

Writing Assessment Symposium (part 2)

12:15-12:45	**Paper 18** **Aynur Ismayilli Karakoc** *University of Bedfordshire*	**What do raters attend to and how are they guided by latent criteria when evaluating integrated reading-writing essays? Evidence from think-aloud protocols**

The current trend in L2 writing assessment has attached great importance to integrated writing assessment tasks which often require incorporating reading into writing. Nevertheless, there is a dearth of studies on understanding the nature of rating the integrated writing assessment tasks, which is a sine qua non for scoring validity. For this reason, investigating rating processes and rater justifications when evaluating integrated writing is important.

This study reports on the raters' cognitive processes when evaluating the writers' performance data based on an analytical rating scale. The test includes two reading texts, and it requires writing an essay. The rating scale is analytic, and it evaluates test-takers' reading and writing integration skills. Integrated essays (N=68) written by EAP students at a New Zealand university were rated by experienced instructors (N=6). Each essay was double-marked. The raters recorded their thinking while rating the essays (in total, N=78). The think-aloud protocols (TAPs) were transcribed and then coded in NVIVO. The data were qualitatively analysed to find patterns. Findings provide insights into the raters' cognitive processes when rating integrated essays. One major theme represents how the raters evaluated the performance data. Another major theme is related to the raters' evaluations of the adequacy or inadequacy of the rating scale. Additionally, the findings shed light on the extent to which the raters were guided by latent or intuitive criteria regardless of the analytical rating scale. The outcomes of this research have implications for developing and refining rating scales and rater training for integrated reading-writing tests.

12:45-13:15	**Paper 19**	**Scoring integrated reading-into-writing**
	Santi Lestari	**performance: Approaches to construct**
	Lancaster University	**operationalisation and their potential impact**

Integrated reading-into-writing tasks, in which test-takers are required to compose a writing response using information presented in (multiple) reading passages usually known as source texts, have been found to elicit a unique construct beyond the sum of its parts (i.e., reading and writing). This unique construct is, to a large extent, due to the requirement for test-takers to synthesise information from the multiple source texts and reformulate it to fulfil the task demand. While this unique construct has contributed to enhancing the authenticity and fairness of reading-into-writing tasks, operationalising this construct into rating scales is not straightforward. At least two approaches to reading-into-writing construct operationalisation into analytic rating scales have been identified: one, where the reading-related aspects of the construct are operationalised into one or more separate criteria within an analytic rating scale, and another, where these aspects are operationalised into one or more specific criteria and also interwoven into other criteria representing general writing and linguistic aspects.

This presentation will report on a study investigating the potential impact of these different approaches to construct operationalisation on rating. The study employed a multistage mixed-methods research design involving psychometric analyses of scores produced by 20 raters scoring a set of reading-into-writing performances using two different analytic rating scales, each representing a different approach, and qualitative analyses of 15 raters' think-aloud protocols and post-rating interview data. The findings suggest that both approaches resulted in quality scoring, evidenced by both psychometric and qualitative evidence. Several differences, however, were observed such as with regards to raters' views of the distinguishability of the different criteria within a rating scale. The qualitative data also further shed light on raters' approaches to and perceptions of source use evaluation in integrated reading-into-writing scoring.

13:15-13:45	**Paper 20**	**Evolution of L2 writing assessment**
	Tony Green	**+**
	University of Bedfordshire	**Wrap-up of the writing assessment symposium**

In this final paper of the Writing Assessment Symposium, I will draw together key themes, placing issues raised at the conference in the context of longer-term developments in the assessment of writing. As we plunge deeper into the digital age, new options for multimodal communication raise interesting questions about what it is that we should assess. At the same time, technology is opening exciting opportunities for monitoring and scoring written products and writing processes. Many of the issues that face us today would be familiar to the pioneers of examinations in modern languages in the nineteenth and early twentieth centuries, but changing values call for new methods and we should work to ensure that our uses of technology align with our educational values. I will conclude with some suggestions, based on what we have learnt from the conference, about how L2 writing assessment may progress from here.

13:45-14:00 **Awards, Farewell and LTF2023**
Best Student Presentation Award: Sponsored by Duolingo
Best Poster Award: Sponsored by TextInspector
Student Travel Awards: Sponsored by UKALTA

14:00-15:30 **Lunch (Subsidised by UKALTA)**
4th Floor Foyer, 35 Berkeley Square

Abstracts of Recorded Paper Presentations

Recording 01

Teachers' language assessment literacy in the digital age: the construct and the affecting factors

Jing Zhang University of Bristol

The importance of teachers' language assessment literacy (LAL) is widely acknowledged. However, due to the multi-faceted (Fulcher, 2012), contextually situated (Xu & Brown, 2016), all-inclusive (Giraldo & Murcia, 2019) and developing (Taylor, 2013) nature of LAL, there still lacks a widely received definition or framework of teachers' LAL (Stabler-Havener, 2018; Giraldo, 2018). In the digital age, as Bennett (2006) contends, it is inevitable to incorporate technology into assessment and the impact technology brings to language assessment (Yu & Zhang, 2017) cannot be ignored. Harding (2018) brings up the concept of technology-aware LAL, resonating with Eyal's (2012) digital assessment literacy, both highlighting the challenges teachers face in the technology-rich environment and the up-to-date competencies they should acquire, which may become more urgent and pertinent under the impact of the Covid-19 pandemic (Tian, Louw & Khan, 2021; Zhang, Yan & Wang, 2021; Zou, Kong & Lee, 2021). Yet, the digital aspect is not considered in most studies that intend to explore the construct of teachers' LAL.

Targeting China's university English teachers, who teach the largest number of adult English learners in the world (Xu, 2017) and undertake challenging assessment responsibilities (Jin, 2018), this study adopts a mixed-method approach, which involves a large-scale survey with follow-up interviews, to explore the construct of teachers' LAL in the digital age. The produced LAL framework includes a clear digital language assessment component and indicates that teachers think they should develop the most competence in classroom-based assessment. The framework also illustrates how specific contextual and personal factors influence the LAL construct, and administrative title, research interest and teaching major turned out to have a significant impact on teachers' perception of what is important within LAL.

Recording 02

What can eye-tracking tell us about teachers' ratings of EFL students' texts?

Ebtesam Abdulhaleem King Saud University
Suhad Sonbul Umm Al-Qura University
Dina El-Dakhs Prince Sultan University

Several recent studies have employed the eye-tracking technique to examine the online rating processes of written texts, but most of this research has either investigated rater-rubric interaction (Winke & Lime, 2015) or raters' reaction to L1 versus L2 texts (Eckstein et al., 2019). We are not aware of any study that looked into the actual grading process. The current mixed-methods study explores the rating behaviors and underlying cognitive processes exhibited by native and non-native raters as they evaluate the written production of L2 learners. Forty English language teachers (20 natives and 20 non-natives) participated in the study. We targeted 12 English texts under two conditions: six 'original' texts (i.e., as written by L1 Arabic EFL learners with identified mistakes) and six 'corrected' versions of the same texts. 60 total mistakes were categorized as relating to grammar (n=20), vocabulary (n=20), or mechanics (n=20). Each participant read 6 texts (three 'original' and three 'corrected') on the computer screen (using the eye-tracker) and was asked to rate them holistically using a 10-point scale. Participants were then interviewed to gauge their rating behavior. Comparing teachers' ratings revealed no significant differences between the natives and non-natives. Quantitative analysis (mixed-effects modelling) of eye-movement data (total reading time or TRT and first-pass reading time or FPRT) was conducted including: Group (native vs. nonnative), Item Category (grammar, vocabulary, and mechanics), and Condition (original vs. corrected). Results indicated that both groups focused less on mechanics and more on grammar and vocabulary during the initial reading (FPRT measure). However, later reading processes (TRT measure) showed that most attention was paid to vocabulary. Natives spent less time rating the passages and showed slower reading behavior for the 'original' passages over their 'corrected' counterparts. We will discuss implications of the quantitative and qualitative findings to the training of L2 raters.

Recording 03

Test-taking mediums as a facet of performance?

Gergely Dávid Eötvös University Budapest

Although long in coming, computer-based testing appears to have finally arrived on the heels of the pandemic, at least in the author’s home country. The speed of the change presented a challenge to computer-based foreign language test producers since they had to diversify into take-at-home (online) test-taking. The subject of this presentation is a performance variable related to the test-taker, originating in their choice of the test-taking medium in a fully digitalised language examination. It is hoped that the research presented will advance our understanding of test-taking mediums and thus contribute towards a theory of performance. Administering both the version taken at a testing centre (CBT) as well as the one at home (IBT) has created a performance variable that is not only observable, but also measurable. The mediums of test-taking constitute construct-irrelevant variance (Messick) and the related variance therefore has to be separated from the construct-relevant variance. If examination results were calculated and reported in raw scores, variance from the mediums would likely affect the construct of foreign language proficiency itself, which would not be a desirable outcome. According to initial raw score observations, test-takers using the IBT (online) medium reached higher scores (means), while those taking the test at a testing centre scored lower, both in German and English at CEFR level B2, on the basis of data from cca. 1300 test-takers in German and cca. 5000 in English. A follow-up analysis with Many-facet Rasch Analysis was used to assess the possibility – and confirming the existence -- of a test-taking medium facet of performance. If scores are calibrated with MFRM, it is possible to make them reflect the language proficiency construct rather than the effect of the medium the test was taken in. Finally, rival interpretations will be discussed.

---------- **multimodal constructs in language assessment** --------

Recording 04

Young language learners’ cognitive processes of digitalised picture-based causal explanation speaking tasks: An eye-tracking study

Wenjun Ding University of Bristol

Recent years have witnessed an increasing use of digitalised multimodal assessment tasks for young language learners (YLLs) globally, while very limited validation studies looked at YLLs' cognitive processes of the multimodal task features under exam conditions. Investigating YLLs’ cognitive processes offers valuable validity evidence of how such a special group of test takers with growing cognitive abilities might process digitalised multimodal features. Meanwhile, the EFL curriculum reform in China and internationally has promoted the use of open-ended questions like why questions and meaning-focused tasks, yet there is a lack of assessment tasks designed to measure such an instructional focus. As such, this study developed digitalised picture-based causal explanation speaking tasks (CESTs) for YLLs and explored their cognitive processes of CESTs via eye-tracking. Ninety-six Chinese primary-school EFL learners (aged from 9 to 12) completed two CESTs in Chinese and English with eye movements recorded, and two English vocabulary size tests. I investigated YLLs' L1 performance and L1 cognitive processes in relation to their grade levels to build a cognitive baseline for using CESTs to assess L2. I further examined YLLs’ L2 cognitive processes in relation to their L1 cognitive processes, L2 performance scores, L2 vocabulary sizes, and grade levels to explore how L2 linguistic resources and age are related to their cognitive processes. It was found that 4th and 6th graders had similar L1 cognitive processes and L1 scores. In the L2 performance, participants with fewer linguistic resources viewed both the content-relevant area and the areas not directly related to the content significantly longer and more frequently. This study revealed the dynamic interactions between the visual and textual stimuli of the CESTs and YLLs’ cognitive ability and L2 proficiency. I will discuss the implications for the future digital language assessment for YLLs.

Recording 05

Assessing the language aptitude of YLs: multimodality and digital tools

Jasenka Čengić University of Zagreb

Foreign language (FL) aptitude has been defined as a set of cognitive abilities that play a major role in both second and FL learning. (Li, 2022). Apart from the MLAT-Elementary (Carroll & Sapon, 2002), which has shown good predictive validity in several studies conducted in different L1 contexts (Kiss & Nikolov, 2005; Tellier & Roehr-Brackin, 2017), there have been no systematic attempts at designing new measures of FL aptitude of young learners (YLs). Furthermore, traditional aptitude measures like the MLAT-E do not cater to multimodality surrounding YLs. During the validation process three studies (N=49; N=207, N=209) were conducted involving YLs ages 6 & 7 in order to establish the final product of the validation: a novel aptitude test. A measure consisting of both visual and auditory input was designed using a natural language (Hungarian) which was at the same time a new language to the participants whose native language is Croatian. Three tasks where designed and delivered using the Gorilla.sc platform. The paired associates task showed neither internal not external validity despite following an age appropriate design. In addition to this, a measure of language analytic ability (LAA) proved to be both internally and externally consistent but showed very low predictivity. The third measure, an auditory alertness (AA) task operationalized as a number learning task, proved to be both internally and externally valid and showed good predictive validity but received a rather unusual classification from the participants. Namely, YLs failed to see the new language in this task because the content of the task were only numbers. The results are discussed considering the emerging (YL) aptitude theory.

Recording 06

Are plan/map-labelling tasks in listening assessments simply testing listening?

Sharon Wong University of Bristol/Chinese University of Hong Kong

Listening test takers' cognitive processes are under-researched but informative regarding the quality of a listening test. By looking into L2 test takers' cognitive processes at real time, this study aims at evaluating the construct validity of plan/map-labelling tasks in IELTS, which requires test takers to relate aural input to visual representation with spatial elements. A mixed-method approach is adopted: Listening scores, visual-spatial ability (VSA) test scores and eye-tracking measurements for statistical analysis; participants' stimulated recall upon viewing their own eye-movement replay as qualitative data. Each of the 40 participants performed 4 tasks with visual-spatial elements (plan/map-labelling tasks) (VS condition) and 4 without (non-VS condition). Results showed that test takers performed significantly better in non-VS tasks, but participants with higher VSA did not perform better than those with lower VSA, even in the VS condition. Their cognitive processes in the two conditions were different. In the VS condition, their response time was significantly longer, and they constantly fixated on previous questions. Data from the stimulated recall suggested that they were not sure about previous answers and therefore they looked back, recalled what they had heard and reconsidered the answers. Given a larger area of interest (AOI) to process and a shorter time span, test takers doing plan/map-labelling tasks were forced to process more information within a shorter period of time and to respond more quickly to one question as the audio moved on to the next. They also reported they were at a high level of stress when they saw this type of questions. The cognitive processes revealed explained participants' lower performance in the VS condition, and have implications on the cognitive validity of this task type and on including visuals in listening assessments. Practitioners should take into account the extra cognitive load involved in tasks with visual-spatial elements when designing listening tests.

Recording 07

Exploring multimodality in L2 assessment through the lenses of eye-tracking: A systematic review

Xin Hu Nanyang Technological University
Vahid Aryadoust Nanyang Technological University

Eye-tracking is an online data collection method for recording participants' gaze behaviors, eye movements, and pupil dilation. Given the capacity of eye-tracking to capture the moment-by-moment processing of written and oral modalities and non-verbal semiotics in an unobtrusive way, this technique has been increasingly applied by second language (L2) researchers to understand the processing, acquisition, and assessment of additional languages. The purpose of this systematic review is to present a synthesis of the use of eye-tracking in L2 research in terms of cognitive mechanisms and eye-tracking measures applied in various modalities. The selection of appropriate and reliable eye-tracking measures that can shed light on the construct of interest is an important element of applied eye-tracking studies, especially since modern eye-tracking technologies can provide researchers with a large amount of gaze data and varied measures of gaze behavior. Following the PRISMA guidelines and using the Scopus database, we identified 111 L2 studies from 17 quartile-1 journals that used eye-tracking. A coding scheme was developed to identify the main themes of each study and extract their methodological practices. We grouped eye-tracking measures into eight main types: fixation, dwell, saccade, skip, regression, pupillometry, blink, and gaze patterns. Furthermore, the first four measure types were differentiated as three scales: temporal, spatial, and count scales. The findings show three main types of cognitive mechanism examined through eye-tracking in various modalities: attention, cognitive processing, and cognitive load. Specifically, attention was predominantly measured via fixation temporal indices, while cognitive processing was frequently measured by using fixation count measures and fixation temporal measures. In addition, the measures adopted to assess cognitive load mainly depended on the task type.

Recording 08

Examining the visual semiotics in a computerized multimodal listening assessment: The role of gaze and mouse-clicking

Yue Qiu Nanyang Technological University
Vahid Aryadoust Nanyang Technological University

There has been a growing interest in the use of computer-facilitated techniques and multimodal texts as inputs in language assessment. As a feature of computerized tests, multimedia provides rich semiotic inputs by using images, sounds, and videos, potentially enhancing authenticity of tests. Studies in human-computer interactions have discussed the multimodal interfaces consisting of two or more combined user input modes, such as speech, gaze, speech, touch, mouse clicking, manual gestures, and body movements.

Using eye-tracking technology involving 77 participants from a major university in Singapore, this research investigated the differences in gaze and mouse-clicking behaviors of L1 and L2 listeners in a commercialized computer-mediated listening test. This test consists of 4 sections and 40 items covering map-labelling, MCQ, table completion, note completion etc. A linear mixed-effects model and repeated measures MANOVA were applied to examine the role of gaze and mouse-clicking behaviors as well as language background and the effect of listening section on listening performance. The results showed significant differences in gaze behaviors and mouse click counts across the four sections, and between L1 and L2. The test is progressively difficult and L1 listeners performed significantly better than their L2 counterparts. The linear mixed-effects model in RStudio revealed that fixation duration and visit counts positively predicted listening performance, while fixation counts, mouse click counts and English as L2 negatively predicted listening scores. These fixed effects have differential predictive power across the four sections and between L1 and L2. Implications for multimodal listening assessment and the application of gaze and mouse-clicking in the multimodal assessment of listening are discussed.

Recording 09

Beyond the Four Skills – Using Modes to Assess Integrated Tasks and Academic Literacy

Chris R Smith University of Edinburgh

There has been significant interest in integrated tasks in recent years, particularly in the sphere of EAP, as these better represent the target language use domain and the complex multimodal nature of language use. However, these integrated tasks are often described and assessed using a four separated skills framework that may not capture the complexity of the tasks. This talk will describe integrated tasks using the CEFR's concepts of modes to go beyond the four skills towards a more complete description of multimodal assessments. One common integrated task in EAP is reading-into-writing, and while the name clearly shows the main skills, there is usually more involved in the task than the simple combination of reading comprehension and written production, with paraphrase, synthesis and evaluation all being possible. This talk will argue that these subskills are better conceptualised separately to reading and writing, and will use the CEFR's concept of "mediation", an umbrella term that includes processing text, as a framework to describe these tasks, and offer suggestions for both formative and summative marking criteria. Another integrated task is the listening-into-speaking seminar, which will also be analysed using the mode of mediation to provide a fuller description of the task. This description will then be extended by adding a framework for describing interaction, drawing on the CEFR and others who have elevated interaction above the mere combination of comprehension and production. The talk will also discuss interaction in writing, which is increasingly important in the digital age. Finally, the ideas of the talk will be interwoven to offer a framework for describing and assessing EAP that goes beyond the four skills, towards a model that has a much wider coverage of the construct and better represents academic language and literacy.

Recording 10

From paper handouts to interactive websites: technology for students' self-assessment in the language learning classroom

Geraldine Bengsch University of York

Technology has the power to profoundly reshape contemporary practices in everyday life (Barton & Lee, 2013). The Covid-19 pandemic has strongly influenced the teaching and learning landscape in Higher Education. While the presence of tablets and laptops may have not been welcomed by every instructor a few years ago, the pandemic has shown that technology online and in face-to-face instruction is here to stay. Handouts are now often distributed in a digital format; however, they are not necessarily adapted for the new medium, and stay in a static form. Online components in language learning are not new, but often remain a contested element of instruction. Technology does not constitute an inherent benefit in instruction (Godwin-Jones, 2014), rather, it is based on implementation into the class where it digital activities have shown to create useful experiences in language proficiency (Alfehaid, 2013).

In this presentation, I reflect on my experience of re-designing my past handouts for a web experience in a beginner's German course at the University of York, UK. Activity handouts were converted into a gamified online experience and published as a website to the internet. Here, HTML, CSS and JavaScript were used to create simple pages with movable elements for fill in the blanks, matching and other exercises. Videos, pictures, and other media can be used for students to engage with. This creates a naturally multimodal experience to support learner autonomy. The aim was to use design and user experience principles to create meaningful learning in and outside the classroom (Krystalli, Panagiotidis, & Arvanitis, 2020; Marden & Herrington, 2020). Handouts, their use in the classroom and their purpose can become the subject of greater scrutiny through the directed focus on instructional design for interactive online games, including re-engagement with the analytical need for the activity, its design, development and implementation into a new medium, and its evaluation through student engagement with it (Fatih, 2016; Stefaniak, 2020).

Recording 11

Exploring the role of listening comprehension in listening-to-summarize tasks: Preliminary Findings

Mikako Nishikawa Nagasaki University
Yukio Horiguchi Kansai University
Guoxing Yu University of Bristol

Success in higher education, regardless of language, requires the ability to incorporate ideas from independent sources. Recently, there has been a growing trend toward assessing the English abilities of Learners of English as a second language (L2) to summarize different sources (e.g., texts, graphs, images, and lectures). Past studies have also shown that the academic success of L2 learners depended on listening skills during class hours (Coakley & Wolvin, 1997; Ferris, 1998; Volgely, 1998). Listen-to-summarize tasks are more complex than listening alone because it requires a combination of skills, such as speaking or writing. To date, fewer empirical studies are related to listen-to-summarize tasks in L2 contexts (Wang & Yu, 2018; Rukthong & Brunfat, 2020). Our study aims to identify factors that could successfully predict the test performance of the listening-to-summarize tasks by comparing participants' processes of listening-to-summarize tasks for writing. In this mixed-methods study, we collected data from eye-tracking data, notetaking strategy questionnaires, and stimulated-recall interviews. In this presentation, we will report our preliminary findings (N=7) based on 1) the number of words in the summary, 2) note-taking strategies questionnaire responses, and 3) the average total fixation duration during listening comprehension. Findings suggest how listening comprehension impacted test-takers from the average fixation duration during listening.

Abstracts of Poster Presentations

A Critical Examination in the General English Proficiency Test (GEPT) Intermediate Writing Test

Li-Chung, Chang University College London

Taiwan aims to make itself a bilingual nation, where both English and Mandarin Chinese are regarded as official languages, in 2030 and has recently started promoting bilingual education. This then highlights the urgent need to develop a reliable and valid test that can adequately measure one's English proficiency level. The General English Proficiency Test (GEPT) is a localized test developed by the Language Training and Testing Center (LTTC) and aims to provide both academic and non-academic institutions a reference of the test taker's English ability. However, despite its prevalence, there is a lack of research examining its validity, with most of its research conducted in-house and not peer-reviewed. This article adopted the argument-based validity approach to examine the validity of its Intermediate Writing Test. After carefully analyzing its test developmental process, its sample test questions, and relevant research, this paper argues that the test is reliable and is impartial to test takers around Taiwan. However, there are still several concerns regarding its validity issue. Although the test can well measure learners' general English ability, it might not be generalizable to English for other purposes, such as business uses. That is, the writing test can only measure one's general English ability, but it is not sufficient to measure English for specific purposes, as LTTC claims. Therefore, LTTC might need to either redefine its test constructs or narrow the aim of the test, so as to strengthen the validity of the Intermediate Writing Test.

Imagined Competencies: A Case Study of Washback in regard to Student Attainment, Teaching Practice, and Language Ideology from Standardised Testing at Public Universities in Southern Mexico

Alexander Black University College London

Aims: This study aimed to assess the impact of assessment systems used in public universities in Oaxaca, Mexico. The TOEFL ITP has long been used to assess student proficiency as a requirement for graduation and funding (UMAR, 2016; CONACYT, 2022). The poor validity and significant washback of this metric have caused critics to appeal for research into alternative testing instruments (Hernandez, 2021; López-Gopar, 2021); further concerns regarding its impact on language attitudes and ideology have also been raised (Sayer, 2012; De Korne, 2021). This study compares washback from the TOEFL ITP and the Cambridge FCE at two Oaxacan universities of comparable size, staff and student profile. The aim of the research is explicitly not to endorse either ETS or Cambridge examination systems, but rather to investigate the unintended negative impact of respective examinations on ELT with a view to exploring shortcomings and alternatives.

This study represents the first cycle of a broader piece of participatory action research on English Language Teaching in Higher Education in Oaxaca, which hopes to attract interest and participation from other teachers and researchers in the region.

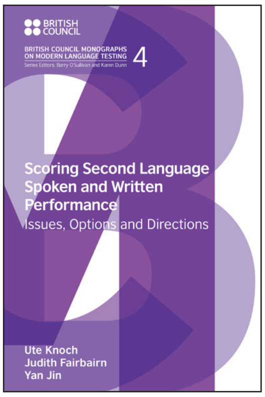
Methods: Research questions centered on the impact of examination systems at two comparable universities on student graduation rates, teaching practices, and language ideology. Secondary data collection was used to analyse successful student pass rates on completion of their program at respective universities. Teachers were interviewed on their teaching practice and interviews transcribed and coded using Otter AI and NVivo, with document analysis of syllabi and teaching materials used for triangulation. Finally, questionnaires were issued to students entering both programs and recent graduates in an attempt to assess the influence of their course of study on language attitudes and ideologies; this was achieved using online multiple-choice questionnaires on Google Forms.

Results: Preliminary results suggest higher pass rates from the group using Cambridge exam systems despite the ostensibly higher cut-off score. Classrooms teaching to Cambridge assessment systems were observed to follow a communicative methodology, with TOEFL ITP encouraging more focus on grammar and reading, following a Grammar Translation approach. Questionnaire results suggest multiple findings of interest including native-speakerist perceptions amongst Cambridge examinees and minoritising attitudes toward indigenous languages from TOEFL ITP sitters.

British Council Monographs on Modern Language Testing

25% DISCOUNT AVAILABLE

Published in co-operation with the British Council, this series offer short books on language testing. Written by well-known language testing scholars from across the world and members of the British Council's Assessment Research Group, these books offer theoretical and practical perspectives to language testing and assessment. They are authored by individuals with considerable academic, teaching and assessment experience, providing the reader with a unique insight into the link between theory and practice. In many cases, the books illustrate their approach with reference to actual test items from the British Council's Aptis language proficiency test service.


BRITISH COUNCIL
BRITISH COUNCIL MONOGRAPHS ON MODERN LANGUAGE TESTING
Series Editors: Barry O'Sullivan and Karen Dunn
4
Scoring Second Language Spoken and Written Performance
Issues, Options and Directions
Ute Knoch
Judith Fairbairn
Yan Jin

Scoring Second Language Performance Assessments

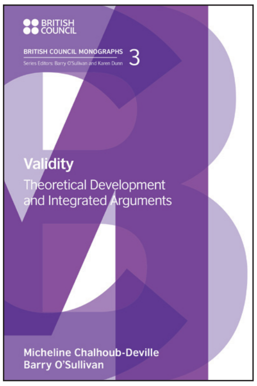
Ute Knoch, Judith Fairbairn and Yan Jin

Because effective speaking and writing skills are considered important in second and foreign language learning contexts, they are often included in performance assessments. However, the scoring of such performances is a complex undertaking and the increasing use of automated scoring systems has added to this complexity in recent years. This timely volume draws together the latest literature on the scoring of second language performances. It focusses on issues relating to rater-mediated assessments and sets out consideration in relation to automated scoring systems, and other technologies, that are increasingly used in the field. This unique volume provides an invaluable introduction to this topic to graduate students, researchers, test developers, other practitioners and teachers.

2021 202pp 234 x 156mm eBook ISBN 9781781799536
pb ISBN 9781781799529 ~~£24.95 / $32.00~~ £18.71 / $24.00

Validity: Theoretical Development and Integrated Arguments

Micheline Chalhoub-Deville and Barry O'Sullivan

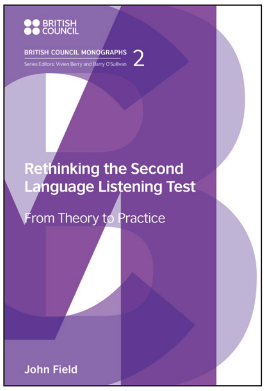

BRITISH COUNCIL
BRITISH COUNCIL MONOGRAPHS
Series Editors: Barry O'Sullivan and Karen Dunn
3
Validity
Theoretical Development and Integrated Arguments
Micheline Chalhoub-Deville
Barry O'Sullivan

This book explores theoretical notions of validity, as well as pragmatic validation practices, and expands the arguments that need to be attended to document quality. The authors examine the need to consider, in addition to psychometric evidence, other critical sources of quality evidence. They promote the concept of impact by design, and envision validity scholarship to attend to consequences at many different levels. Concomitant with this attention to consequences are considerations of stakeholders and the tailoring of communication to engage intended groups, yielding a more convincing validity argument. The authors call on professionals in the field to publish case studies that showcase localised validity arguments in practice. Local case studies represent critical endeavours to illustrate how evidence and arguments are pulled together to support the quality of a testing programme and all that it entails.

2020 216pp 234 x 156mm Illus. eBook ISBN 9781781799918
pb ISBN 9781781799901 ~~£24.95 / $32.00~~ £18.71 / $24.00

Rethinking the Second Language Listening Test: From Theory to Practice

John Field


BRITISH COUNCIL
BRITISH COUNCIL MONOGRAPHS
Series Editors: Vivien Berry and Barry O'Sullivan
2
Rethinking the Second Language Listening Test
From Theory to Practice
John Field

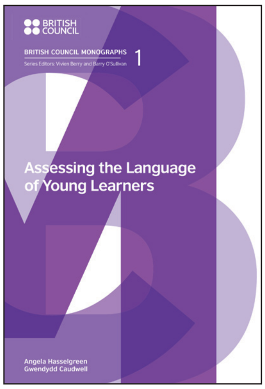
"A useful and timely contribution to the relatively sparse literature on listening test design. Field draws together various elements of his recent empirical and theoretical work to present to the reader his well-known listening process model, and to elaborate on the implications of this model for listening test design. This book will provide a welcome bridge between the more scholarly aspects of Field's work, and the very practical needs of those working in test development, whether in testing organisations or in the language classroom."

Luke Harding, Dept of Applied Linguistics and English Language, Lancaster University, UK

2019 168pp 234 x 156mm eBook ISBN 9781781797167
pb ISBN 9781781797150 ~~£22.95 / $29.95~~ £17.21 / $22.46

Assessing the Language of Young Learners

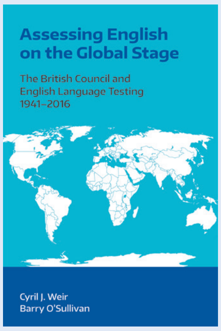
Angela Hasselgreen and Gwendydd Caudwell


BRITISH COUNCIL
BRITISH COUNCIL MONOGRAPHS
Series Editors: Vivien Berry and Barry O'Sullivan
1
Assessing the Language of Young Learners
Angela Hasselgreen
Gwendydd Caudwell

This volume offers new insights into the assessment of the language of Young Learners (YLs) from 5 to 17 years. The authors addresses fundamental issues and consider the ability of different age groups to perform in a second or foreign language, presenting principles of formative assessment and testing in the light of linguistic, cognitive and social development. Focussing on testing a range of 'skills', theoretical models of performance are introduced, followed by a practical analysis of approaches to the testing of each skill. The authors conclude by summing up developmental characteristics of each age group, and their implications for language testing.

Researchers within the field of teaching and assessing the language of young learners are offered scope for further investigation, while practitioners will find support in their day-to-day work with YLs.

2016 172pp 234 x 156mm Illus. eBook ISBN 9781781794760
pb ISBN 9781781794708 ~~£22.95 / $29.95~~ £17.21 / $22.46


Assessing English on the Global Stage
The British Council and English Language Testing 1941–2016
Cyril J. Weir
Barry O'Sullivan

Assessing English on the Global Stage: The British Council and English Language Testing, 1941-2016

Cyril J. Weir and Barry O'Sullivan

This book tells the story of the British Council's seventy-five year involvement in the field of English language testing. It explores the role of the British Council in spreading British influence around the world through the export of British English language examinations and British expertise in language testing.

2017 392pp 234 x 156mm Illus.
eBook ISBN 9781781796016
pb ISBN 9781781794920
~~£38.95/$49.95~~ £29.21/$37.46

Order any book from **equinoxpub.com** and claim a 25% discount using the code **britishcouncil**

The publisher reserves the right to revise prices at any time

equinox
www.equinoxpub.com

TRINITY
COLLEGE LONDON

The international exam board for English language and teaching qualifications

TESOL

Teaching English to Speakers of Other Languages

GESE

Graded Examinations in Spoken English

ISE

Integrated Skills in English

Trinity Stars

Young Performers in English

A full member of ALTE Association of Language Testers in Europe

trinitycollege.com/English

duolingo english test

englishtest.duolingo.com/research

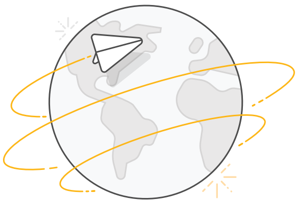
the future of testing. here today.

Built on the latest assessment science

The Duolingo English Test is a computer adaptive test powered by rigorous research and AI. Results are highly correlated with other assessments, such as the TOEFL and the IELTS.

Protected by innovative test security

Industry-leading security protocols, individual test proctoring, and computer adaptive technology help prevent fraud and cheating and ensure results you can trust.



Expands applicant pools

Tap into a diverse pool of candidates from 210+ countries and territories of origin, who have taken the Duolingo English Test because of its radical accessibility.

BRITISH COUNCIL

Assessment Research Group

Supporting research and learning

The British Council supports research in a number of ways. We offer research grants, publish reports and undertake research projects. Apply for a grant or access a publication on the links below.

Enabling innovative research through awards and grants

Assessment Research Awards and Grants (ARAGs): Available for research students and more experienced researchers, these awards and grants recognise achievement and innovation within the field of language assessment.

www.britishcouncil.org/exam/aptis/research/grants-and-awards

IELTS Joint-funded Research Programme: The IELTS Partners provide funding for educational institutions and researchers to undertake applied research projects in relation to IELTS.

www.ielts.org/for-researchers/research-proposals

English Assessment Research Grant Scheme: This jointly-funded NEEA-British Council scheme provides funding to researchers inside and outside of China to conduct research on English language assessment.

www.britishcouncil.cn/en/exams/cse-research

Reading into Research Grants (RiRGs): MetaMetrics and the British Council invite applications for research into our understanding of the construct of EFL reading comprehension and reading comprehension assessment.

www.britishcouncil.org/exam/aptis/research/grants-and-awards/research-into-reading

Open access to our research publications online

Technical Reports: Test development and validation studies related to the Aptis test system.

Assessment Research Awards and Grants Reports and Non-Technical Summaries: Projects carried out by external researchers that have been funded through our awards scheme.

British Council Validation Series: Studies in collaboration with external researchers to target areas of importance for Aptis and for language assessment generally.

British Council Perspectives on English Language Policy and Education: This series sets out the British Council's approach to a range of issues around the English language in learning systems around the world.

www.britishcouncil.org/exam/aptis/research/publications

Books by testing experts

British Council Monographs on Modern Language Testing: Published in collaboration with Equinox Publishing, these books offer both a theoretical and a practical perspective to language testing and assessment.

www.britishcouncil.org/exam/aptis/research/publications/monograph-series

www.britishcouncil.org/exam/aptis/research

UKALTA

UK Association for Language Testing and Assessment

BRITISH COUNCIL

LTF2023

24-26 November 2023

Hosted by

British Council
1 Redman Place
Stratford, London
E20 1JQ

BRITISH COUNCIL

Potential Themes

1. Assessing plurilingual competence
2. Testing communicative competence: Beyond language
3. Innovative approaches in language testing
4. Validity, fairness and transparency with technology in testing
5. Comprehensibility: Challenges for assessment
6. Customised learning and assessment
7. Your suggestions…

➢ **Use the QR code to vote** →

UKALTA
UK Association for Language Testing and Assessment

BRITISH COUNCIL

Thank you for your input, we look forward to seeing you in Stratford, London in 2023!

BRITISH
COUNC
Our work in arts and culture, education and English creates opportunities for young people across the world to come together to address the climate emergency.
Empowering young people
Engaging our global network
The Prince of Wales Suite

University of BRISTOL

School of Education

35 Berkeley Square

Bristol, BS8 1JA, UK

bristol.ac.uk/education