LTF 2025
LANGUAGE TESTING FORUM
SOUTHAMPTON
21-23 NOV

CELEBRATING 45 YEARS LTF

CELEBRATING 10 YEARS UKALTA

LANGUAGE TESTING FORUM 2025

Interdisciplinary trends and challenges in language assessment

University of Southampton | LATU Language Assessment and Testing Unit | UKALTA UK ASSOCIATION FOR LANGUAGE TESTING AND ASSESSMENT

CONTENTS

PRESIDENT'S WELCOME

As President of the UK Association for Language Testing and Assessment (UKALTA), I'm pleased to welcome you to the 45th Language Testing Forum (LTF) – our annual conference as a professional association.

We are most grateful to Ying Zheng and the entire LATU team for their hard work in planning and delivering this year's event on the university campus in Southampton with the conference theme of 'Interdisciplinary Trends and Challenges in Language Assessment'. This topic is both timely and relevant, highlighting the growing need for interdisciplinary approaches to address ethical considerations in language testing and assessment, including the challenges posed by rapidly evolving technological advancements and broader societal Issues.

As part of UKALTA's growing Awards and Sponsorship Programme, we are once again supporting one or more graduate students to attend LTF through our Student Awards (now up to £500 per person, subject to need and budget available). The UKALTA/Hornby Scholarship award is being offered for the second year running to support an international ELT scholar undertaking the MEd TESOL at the University of Exeter so they can attend LTF as part of their UK study experience. For 2025 we have introduced a new award to encourage a first-time attendee at LTF. Please do warmly welcome students and any first-time delegates so they quickly feel at home within our professional community.

The large number of high-quality abstracts submitted for LTF means that we are organising some parallel sessions this year to accommodate as many presenters as possible. The programme includes 25 papers, 14 posters and an exciting opening plenary from invited speaker, Prof Denis Drieghe. The stimulating and topical programme includes the usual Saturday evening dinner which this year is being very generously sponsored, enabling us to celebrate two important anniversaries: the 45th anniversary of LTF, and the 10th anniversary of UKALTA. Without the support of our wonderfully generous sponsors, LTF would struggle to exist – so please do thank them if you have the chance!

Before the dinner on Saturday evening, we shall hold the UKALTA Annual General Meeting (AGM) as usual – this year accompanied by a glass of something celebratory! All are welcome to attend the AGM whether or not you are a member of UKALTA, so please do come along. The AGM is a crucial component of our life together as a responsible and effective learned society – a chance for the Executive Committee to report back to the Membership and for members to discuss and decide on important issues going forward. Thank you in advance for your commitment to this key part of the conference programme.

In this my final year as UKALTA President, I'd like to express my deep personal gratitude for all your support over the past 5+ years, and I look forward to a stimulating, productive and enjoyable weekend together.

Best wishes,

Prof. Lynda Taylor
Professor in Language Testing and Assessment, CRELLA

A MESSAGE FROM THE CHAIR

A warm welcome to the Language Testing Forum (LTF) 2025 at the University of Southampton (21–23 November), hosted by the Language Assessment & Testing Unit (LATU) in the Department of Languages, Cultures & Linguistics. LATU's mission, to provide a forum for knowledge generation, dissemination and exchange in language assessment and testing and to build new collaborations, is well aligned with the aims of this conference.

LTF was last in Southampton in November 2014, hard to believe that was 11 years ago! It was right here in this very venue that the idea of starting a UK language testing association was first brought up by the attendees. This year, we're delighted to be celebrating two milestones: 45 years of the Language Testing Forum and 10 years of the UK Association of Language Testing & Assessment (UKALTA).

The theme of this conference is "Interdisciplinary trends and challenges in language assessment". We open with Professor Denis Drieghe's keynote, Reading in Different Languages, followed by exciting parallel papers (an LTF first!), posters, and plenty of time to connect. LTF is proudly collegial and especially welcoming to first-time attendees and PGRs. Please introduce yourselves, ask questions, and join the conversations across sessions and breaks.

Thank you to our presenters, session chairs, student helpers, and sponsors. We wish you all an engaging, friendly and fruitful LTF 2025 in Southampton. May you leave with new ideas and new experiences.

Warmest regards,

Prof. Ying Zheng
Chair, LTF 2025 Organising Team

CYRIL WEIR LECTURE

In this keynote talk, I will introduce the research field of eye movements during reading before focusing on two large projects of cross-linguistic research, including reading in English as a second language.

Eye-tracking is widely regarded as the methodological gold standard for studying cognitive processing during reading. Readers sample information from a line of text through a series of fixations, during which the eyes are relatively still and extract information, and saccades, which are rapid, jerky movements that shift the point of fixation to bring new information into the centre of the visual field, where visual acuity is highest (the fovea). Extensive research has demonstrated that various linguistic factors influence both the duration and location of fixations (see Rayner, 2009, for a review). For example, high-frequency words (e.g., apple) tend to receive shorter fixation durations than less common words (e.g., inlet). Likewise, short words and words that are predictable from the preceding context receive fewer and shorter fixations than long or unpredictable words, reflecting the processing ease associated with those words.

In typical reading experiments, participants read silently from a computer screen while an eye-tracker continuously records their eye positions without disrupting natural reading behaviour. This high ecological validity has contributed to eye-tracking becoming the dominant methodology in reading research, enabling researchers to capture fine-grained details of cognitive processing.

About the speaker

Denis Drieghe received his PhD in Experimental Psychology from Ghent University in Belgium working with Marc Brysbaert. He obtained consecutive positions as a Postdoctoral Research Fellow for a total of 5 years. During his time in Ghent, he also obtained multiple grants to go abroad for research visits. He spent a total of 2.5 years at the University of Massachusetts in Amherst, MA (USA) working with Keith Rayner and Sandy Pollatsek, after which he was a visiting research fellow at the University of Southampton (UK). In 2010 he took on a lectureship at the School of Psychology in Southampton where he is currently a Professor of Experimental Psychology and the Deputy Head of the School (Research). His research can be situated in the field of eye movements during reading. He has examined reading in multiple languages (English, Dutch, Finnish, Chinese, Arabic, Brazilian Portuguese and Hindi), and has directly compared reading in different languages both between native speakers and within bilinguals. His research has been funded by the Fund for Scientific Research Flanders, the Experimental Psychology Society, the Australian Research Council and the Leverhulme Trust.

THANK YOU
TO OUR SPONSORS

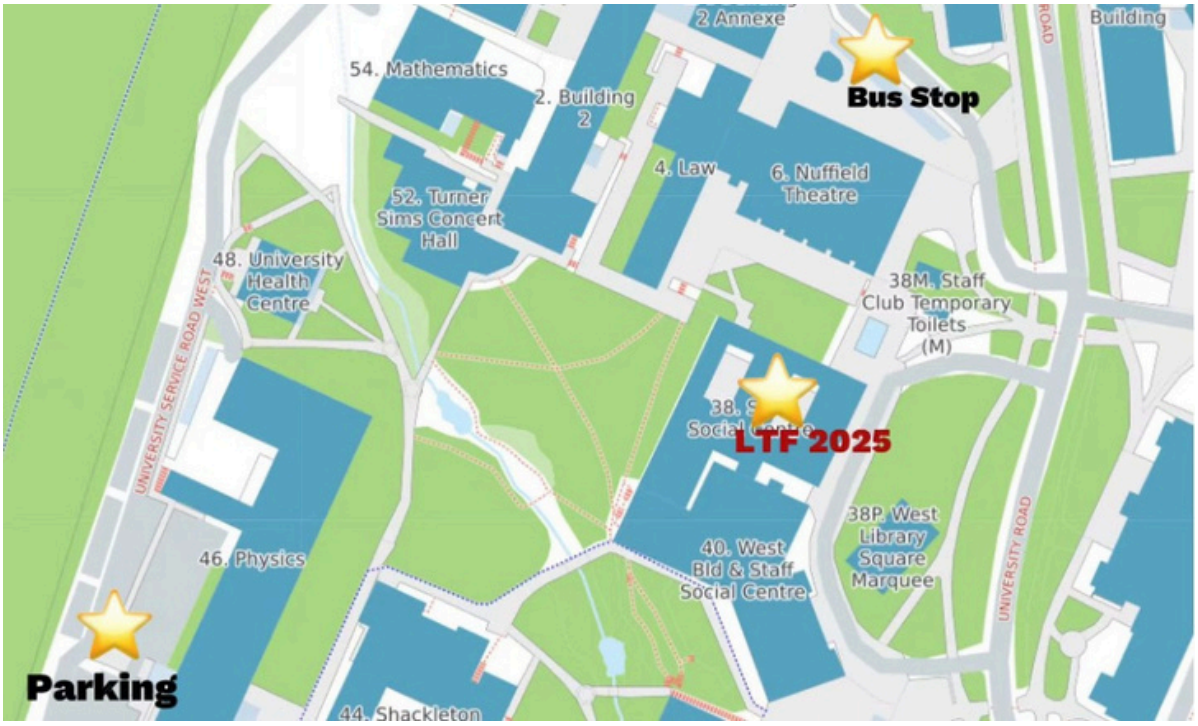
WIFI OPTIONS

The conference venue provides two **Wi-Fi** options:

Eduroam: Available for participants from universities and institutions that support Eduroam. Please log in using your institutional credentials.

WiFi Guest: For participants without Eduroam access. Simply select WiFi Guest from the available networks and complete the quick registration to get connected.

CONFERENCE VENUE



Hartley Suite (Building 38)
University of Southampton
Highfield Campus
University Road, Southampton
SO17 1BJ

SCHEDULE

Friday 21 November 2025

17:00-17:45	**Registration**
17:45-18:00	Welcome & Introduction (Prof. James Ryan, Head of School of Humanities, University of Southampton; Prof. Lynda Taylor, President of UKALTA)
18:00-19:00	Cyril Weir Lecture Speaker: Denis Drieghe (Professor of Experimental Psychology, University of Southampton) Reading in Different Languages
19:00-20:00	Drinks Reception (Sponsored by Oxford University Press)

Saturday 22 November 2025

8:30-9:00	Arrivals & Registration

Garden Court

Chair: John Pill

9:00-9:25	**Paper 1: Developing Automated L2 Pronunciation Assessment with Construct-Aligned ML Models** Masha Kostromitina, Ben Naismith, Geoff LaFlair, Kevin Yancey, Danwei Cai
9:30-9:55	**Paper 2: Operationalising Intelligibility in CEFR-CV via Adaptive Comparative Judgment with Think-Aloud** Jingwen Wang
10:00-10:25	**Paper 3: Prototyping Oral Assessment Design via Multimodal CA: Student Interactions Pre- & Post-Intervention** Katherina Walper, Philipp Siepmann
10:30-11:00	Tea/Coffee Break (Sponsored by Cambridge University Press & Assessment)

	Garden Court	Hartley Suite
	Chair: Özgür Şahan	**Chair: Duygu Candarli**
11:00-11:25		**Paper 7: Supporting L2 Reading Comprehension by Using Artificial Intelligence and Diagnostic-Dynamic Assessment** Ari Huhta, Dmitri Leontjev
11:30-11:55	**Paper 5: Leveraging Generative AI to Establish Linguistic Features Associated with Spoken English Proficiency Levels**	

	Johnathan Jones, Leda Lampropoulou	
12:00-12:25	**Paper 6: Investigating Rater Reliability and Cultural Bias in AI-mediated Speaking Assessment** Yasin Karatay, Leyla Karatay, Jing Xu	**Paper 9: Evaluating Chatbot Authenticity in Simulations of Spoken Interaction: Demonstrating the Utility of Corpus-based Methods for Development and Validation** Dana Gablasova, Luke Harding, Vaclav Brezina, Emil T. Hazelhurst, Barry O'Sullivan, Richard Spiby
12:30-12:50	1-minute poster presentations	
12:50-13:00	Group photo	
13:00-14:00	Lunch & posters (Sponsored by Pearson)	

Garden Court

Chair: Gemma Bellhouse

14:00-14:25	**Paper 10: Effects of Script Presentation Mode on Raters' Scoring Outcomes and Cognitive Processes: An Eye-Tracking Study** Daniel Yu-Sheng Chang
14:30-14:55	**Paper 11: Understanding Cognitive Processes Underlying Multiple-text Items Through Eye-Tracking and Verbal Protocol Methods** Aylin Unaldi
15:00-15:30	Tea/Coffee Break (Sponsored by Cambridge University Press & Assessment)

Chair: Sathena Chan

15:30-15:55	**Paper 12: Teacher Literacy in IELTS Writing Assessment: Applying Task 2 Band Descriptors in Formative Feedback** Yasaman Ibayeva
16:00-16:25	**Paper 13: Online EMI Interaction in University Classes: Developing Evidence-Based Resources** Fumiyo Nakatsuhara, Chihiro Inoue, Katherine Halley, Akiko Kiyota, Ridhwan Abdullah, Hasif Jazila, Amilin Nordin, Yasuyo Sawaki
16:30-17:45	**UKALTA AGM** UKALTA 10th and LTF 45th Anniversary Drinks
18:30-21:00	**Dinner** (Subsidised by Duolingo) Southampton Harbour Hotel – The Needles 5 Maritime Walk, Ocean Village Southampton SO14 3QT (Complimentary Parking: Ocean Village Multi-Storey Car Park)

Sunday 23 November 2025

8:30-9:00	Arrivals

	Garden Court	Hartley Suite
	Chair: Leda Lampropoulou	**Chair: Yasin Karatay**
9:00-9:25	**Paper 14: Documenting Validity: A Review of Test Manuals for High-Stakes English Language Tests** Ben Naismith, Ramsey Cardwell, Masha Kostromitina	**Paper 17: Optimizing Computer Adaptive Test Design Through Simulation: Investigating the Oxford Placement Test** Nathaniel Owen
9:30-9:55	**Paper 15: Assessing Metapragmatic Awareness for Intercultural Communication: Harnessing Interdisciplinary Power in Language Assessment** Shishi Zhang	**Paper 18: Human-AI Collaboration in Language Testing: Equity and Validity** Ramsey Cardwell, Alina von Davier, Jill Burstein
10:00-10:25	**Paper 16: Applying the Lessons from Second Language Test Validity Frameworks to Tests of English as a First Language** Nicola Latimer, Fumiyo Nakatsuhara, Sathena Chan, Pauline Madella, Samantha Orciel, Lydia Ridding	**Paper 19: Targeting the Top: Evaluating Differentiation among High-Performing Candidates in Items Designed to Stretch and Challenge** Ana Ulicheva, Sarah R. Hughes
10:30-11:00	Tea/Coffee Break (Sponsored by Cambridge University Press & Assessment)	
	Chair: Nathanial Owen	**Chair: Shishi Zhang**
11:00-11:25	**Paper 20: Beyond the Lab: A Virtual Desktop Data Collection Method for Writing Process Research** Michelle Czajkowski	**Paper 23: At the Intersection of Sociolinguistics and Language Testing: Comprehensibility, Pronunciation Features and English as a Lingua Franca** Sheryl Cooke, Karen Dunn
11:30-11:55	**Paper 21: Can Grammatical Metaphor Be Used as a Marker of Writing Complexity in L2 Assessment?** Nicholas Glasson, Andrew Kitney	**Paper 24: Test Anxiety Revisited: A Comparative Study of At-home vs In-centre High Stakes Tests** Ruolin Hu
12:00-12:25	**Paper 22: AI-generated Voices and Accent Diversity: Towards Fairer Listening Assessment** Leyla Karatay, Nicholas Glasson, Andrew Mullooly	**Paper 25: Phraseological Knowledge in a Dialogic Speaking Task across Assessed Levels of Proficiency** Parvaneh Tavakoli, Takumi Uchihara, Svetlana Mazhurnaya, Phil Smyth
12:30-13:00	Awards, Farewell & LTF2026	
13:00	Packed lunch (Subsidised by Kaplan)	

POSTER PRESENTATIONS

1	**Investigating Cognitive Process in Duolingo English Test Listening Section: An Eye Tracking Approach** Jia Li and Yuxuan Yang, University of Southampton
2	**Comparing Overall Score Distributions in the Evaluation of an Automated Scoring System** Edmund Jones, Trevor Breakspear and Jing Xu, Cambridge University Press & Assessment
3	**“Intelligent Enough to Assess Me?”: Exploring the Intelligence and Effectiveness of AI Feedback in Speaking Language Assessment** Yuchen Xing, University of Southampton
4	**Speaking On-topic: An Enhancement to the Socio-cognitive Framework and Raters’ Perceptions About Feelings** Judith Fairbairn, University College London
5	**PFL proficiency assessment: Factors influencing learner performance in different cloze test formats** Clara Setas, Universidade do Minho (CEHUM)
6	**The Impact of Automated Writing Feedback on Self-Regulated Writing Strategies and L2 Writing Performance among Female EFL Learners at Saudi University** Maishael Alsanad, University of Southampton
7	**Mexican and Brazilian High School EFL teachers´ perspectives: National English Policy and its impact on classroom assessment** Elsa Fernanda Gonzalez, Autonomous University of Tamaulipas; Gladys Quevedo-Camargo, University of Brasilia
8	**Approaches to ‘Testing the Untestable’: A Year of Classroom-based Inquiry into Diverse Student Responses in Upper-secondary English as a Foreign Language (EFL) LiteratureAssessment** Lynn Williams Leppich, Bern University of Teacher Education
9	**A Journey into the Teaching and Assessment of Speaking in Portuguese Classrooms** Margarida Pato, CRELLA, University of Bedfordshire
10	**A Longitudinal Case Study Tracking Undergraduate EMI Learners’ Potential Second Language Speaking Development in a Higher Education Context in Japan** Jamie Lesley, CRELLA, University of Bedfordshire
11	**Towards Inclusive Language Assessment: Enhancing Equity and Diversity in Test Design and Practice** Sarvinoz Umarova, Urgench State University

ABSTRACTS

Paper 1: Developing Automated L2 Pronunciation Assessment with Construct-Aligned ML Models

Masha Kostromitina, Ben Naismith, Geoff LaFlair, Kevin Yancey, Danwei Cai, Duolingo

Automated speech evaluation systems are increasingly used in language testing due to the need for efficient, large-scale assessment of L2 proficiency and documented challenges of using human raters to assess L2 pronunciation such as shared-L1 bias (Evanini & Zechner, 2019; Harding, 2012). However, many existing systems are opaque in how they operationalize pronunciation, and these operationalizations are often misaligned with modern conceptualizations of L2 pronunciation. This study evaluates a machine learning model developed to predict human ratings of L2 pronunciation that are explicitly grounded in theoretically and empirically motivated pronunciation constructs.

We developed an automated pronunciation scoring system trained on 2,624 learner speech samples rated by expert human raters. Raters used CEFR-based rubrics that emphasized core pronunciation constructs: phonological control, segmental articulation, and prosody. These rubrics focused on intelligibility and comprehensibility rather than native-likeness, following research that supports more inclusive standards for L2 pronunciation (Jenkins, 2006; Levis, 2020). The model incorporated multi-level acoustic and linguistic features, including phoneme articulation, vowel duration, intonation contours, and pause distribution.

The model achieved a strong correlation with human ratings (Spearman's ρ = 0.82), closely approximating human–human agreement (ρ = 0.86) and outperforming baseline and commercial systems. It also showed robust discrimination between adjacent CEFR levels, particularly at B1–B2 and B2–C1 transitions. Additionally, model scores correlated with university stakeholders' judgments of comprehensibility (r = 0.67) and academic acceptability (r = 0.70), indicating that it captured pronunciation features relevant to real-world communicative contexts.

This study contributes to construct validity in pronunciation assessment by demonstrating that automated scoring can reflect theoretical and pedagogical understandings of L2 pronunciation. The findings have implications for the development of fairer and more meaningful pronunciation assessments, particularly in high-stakes testing contexts.

Paper 2: Operationalising Intelligibility in CEFR-CV via Adaptive Comparative Judgment with Think-Aloud

Jingwen Wang, University of Southampton

The constructs of intelligibility (listeners' actual understanding) and comprehensibility (perceived ease of understanding) are central to L2 pronunciation research, reflecting a shift from native-speaker norms to recognising English as a lingua franca (ELF). Despite growing interest, these constructs remain underdefined. The 2020 Companion Volume to the CEFR (CEFR-CV) redefines intelligibility by merging it with comprehensibility, yet practical tools for assessment remain limited, hindering both research and pedagogy.

This study explores the use of Adaptive Comparative Judgment (ACJ) to assess CEFR-CV intelligibility. In ACJ, judges compare paired speech samples and select the one better meeting a holistic criterion. To uncover the linguistic features guiding these judgements, Think Aloud Protocols (TAPs) were used. These insights informed the development of empirically grounded constructs aligned with the CEFR-CV. Speech samples were drawn from both controlled and spontaneous tasks. In the controlled task, 12 Mandarin L1 judges assessed 30 Mandarin-speaking L2 English learners; in the spontaneous task, 12 mixed-L1 judges evaluated 30 learners from diverse

linguistic backgrounds. The integration of ACJ scores and TAP data provided strong evidence of ACJ's effectiveness. TAPs generated 1,526 and 917 coded features from the respective tasks.

Findings show that CEFR-CV intelligibility includes not only pronunciation (segmentals, suprasegmentals) but also fluency, grammar, discourse, and acceptability. For spontaneous speech, accentedness and lexis also emerge as key. This study advances the use of ACJ in pronunciation assessment and offers a refined operationalisation of CEFR-CV intelligibility, with implications for language teaching and testing in ELF contexts.

Paper 3: Prototyping Oral Assessment Design via Multimodal CA: Student Interactions Pre- & Post-intervention

Katherina Walper, Newcastle University; Philipp Siepmann, Leibniz University of Hannover

This study introduces a novel approach to prototyping a task-based oral assessment applying multimodal Conversation Analysis (CA) (SSJ, 1974; Mondada, 2016) to pre- and post-intervention interactions. Conducted in 11th-grade EFL classrooms in North Rhine-Westphalia, Germany, it compares a traditional exam (reading-to-speaking, text analysis) with a redesigned task (listening-to-speaking, cooperative planning) developed over four years. The aim was to determine whether the revised task elicited more authentic, naturalistic interaction to better assess students' interactional competence (IC). Using multimodal CA, the study analysed interactional practices (gaze, gesture, turn-taking, and peer assessment) in 4.85 hours of video recordings (24 students), transcribed using Jefferson and Mondada systems. Findings reveal that the pre-intervention design primarily generated monologic sequences. Participants oriented towards the examiner rather than peers, with limited uptake or expansion of prior turns. Contributions were largely non-contingent (students introduced new ideas without building/assessing each other's talk. Conversely, the post-intervention fostered mutual orientation with frequent overlapping talk, contingent assessments, and collaborative negotiation. Embodied practices such as gaze alignment, deictic gestures, and responsive turn design underscored participants' interactional engagement and co-construction of meaning—central to IC.

This work advances understanding of CA's application in interactional task analysis (Lam et al., 2023; Seedhouse & Satar, 2021), showing how multimodal CA can inform language test design and validation. It also proposes a transferable design principle to support authentic, interaction-focused assessments in institutional contexts.

Paper 4: Validating GPT-4 Scoring of Writing and Speaking: Human–AI Alignment and Fairness in High-Stakes Language Assessment

Usman Tahir, University of Oxford, also affiliated with Oxford University Press

As large language models (LLMs) like GPT-4 become increasingly integrated into educational technologies, their use in high-stakes assessment raises urgent questions about scoring validity, reliability, and fairness. This study evaluates whether GPT-4 can generate scores that align with trained human raters in a formal language testing context, using a validation framework grounded in scoring theory and fairness principles. Drawing on 5,000 anonymised Speaking and Writing responses from the Oxford Test of English, a globally administered, CEFR-aligned proficiency test certified by the University of Oxford, each previously scored by trained human assessors, we prompt GPT-4 to assign rubric-based scores and explanations using structured templates.

Following the argument-based approach to validation (Kane, 2006) and methodological standards in automated scoring (Yan et al., 2020), we assess human–AI alignment using Pearson correlation, Quadratic Weighted Kappa (QWK), ±1 band agreement, mean and standard deviation of score differences, and Proportional Reduction in Mean Squared Error (PRMSE). To examine fairness, we

model discrepancies between GPT-4 and human scores as a function of demographic features suchas gender, first language, and region. A qualitative analysis of divergent cases explores construct-level judgement differences, particularly where GPT-4 departs from human raters on task relevance, coherence, or linguistic complexity.

Preliminary analyses indicate strong alignment between GPT-4 and human scores, with agreement levels comparable to those observed between trained raters. However, variation across demographic subgroups suggests the need for continued scrutiny of fairness in machine scoring, particularly in certification contexts. This study extends current work on human–AI alignment by providing empirical evidence on the reliability and fairness of GPT-4 scoring. It introduces a scalable, rubric-based evaluation method and demonstrates how psychometric and demographic analysis can inform responsible AI use. Grounded in a theory-driven validation model, the research offers a replicable approach for assessing the viability of generative models in formal testing environments.

Paper 5: Leveraging Generative AI to Establish Linguistic Features Associated with Spoken English Proficiency Levels

Johnathan Jones, CRELLA, University of Bedfordshire; Leda Lampropoulou, LanguageCert

Despite the crucial role of language frameworks in high-stakes settings (e.g., visa, immigration, and university entrance), there is relatively little work done on level-specific linguistic features associated with critical speech dimensions such as fluency, intelligibility, and comprehensibility (Tavakoli et al., 2020), and even less understanding of how generative AI can enhance assessments of spoken language proficiency. The present study explores the implementation of generative AI to define level-specific linguistic features associated with second language (L2) assessment frameworks, exemplified by the Canadian Language Benchmarks (CLB).

Using 22 publicly available oral performance samples from the Centre for Canadian Language Benchmarks, the study developed an automated pipeline combining automatic speech recognition, GPT-4-driven conversational turn segmentation, and custom fluency measurement scripts. This automated approach significantly streamlines speech analysis and scoring, traditionally labour-intensive tasks requiring human raters. Spearman correlations assessed the alignment between AI-derived speech measures and established CLB proficiency levels, while MANOVAs evaluated statistically significant distinctions among proficiency categories.

An independent validation dataset consisting of 40 expert-rated speech samples further examined the generalisability and reliability of the AI-driven approach. Results explore potential relationships between generative AI-generated speech indices and CLB proficiency levels, informing the use of AI in establishing spoken proficiency measures. This presentation contributes to the ongoing discussion around the application of AI in assessment and teaching contexts, exploring generative AI's potential to support consistent, evidence-based language testing and curriculum development.

Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal, 104*(1), 169-191.

Paper 6: Investigating Rater Reliability and Cultural Bias in AI-mediated Speaking Assessment

Yasin Karatay, Leyla Karatay, Jing Xu, Cambridge University Press & Assessment

Recent advances in conversational AI have enabled systems to mimic sociopragmatic behaviors such as politeness formulas and culturally appropriate language use (Godwin-Jones, 2024). However, given the complex interplay among the social, cultural, and technological factors in AI-mediated interactions and stakeholders' divergent expectations of the pragmatic norms of such interactions (Brandt & Hazel, 2025), it is important to examine rater reliability and bias when candidates from

different cultural backgrounds engage in AI-mediated speaking assessment.

To address this, we developed a prototype spoken dialog system (SDS) to mimic the role of a human interlocutor in a computer-based speaking task, combining rule-based mechanisms with a large language model (GPT-4o). Using a mixed-method design, we explored how the prototype SDS can elicit meaningful interactive oral performances across candidates from different cultural groups. The study posed two research questions:
(1) How reliable are human rater scores on SDS-mediated performances, and is there evidence of cultural bias?
(2) Are the candidate-computer interactions equally effective across different cultural groups based on rater perceptions?

Ninety non-native-English-speaking participants from four different countries (i.e., China, Japan, Germany, and Netherlands) completed the speaking task, and their oral responses elicited by the SDS were assessed by four trained raters using a modified IELTS mark scheme. Results are presented in two main areas. First, multi-faceted Rasch measurement analysis is performed on rater scores to investigate score reliability and potential systematic bias in the ratings across cultural groups. Second, thematic analysis of the retrospective interviews will explore raters' perceptions of the task, focusing on AI invariance in interacting with different cultural groups. The findings will provide insights into fairness in using conversational AI for speaking assessment.

Paper 7: Supporting L2 Reading Comprehension by Using Artificial Intelligence and Diagnostic-dynamic Assessment

Ari Huhta, Dmitri Leontjev, University of Jyväskylä

We report on an assessment project employing an AI system to support L2 English reading development among upper secondary school learners in Finland. The project integrates diagnostic language assessment (DiagA) and dynamic assessment (DA) into a coherent framework. The former is an established approach for analysing learners' strengths and weaknesses, with constructs informed by SLA research (Alderson 2005; Alderson et al., 2015). The latter follows Vygotsky's (2012) position that the inclusion of mediation (i.e., support to learners) expands the diagnosis to include abilities that are still emerging (Poehner & Lantolf, 2023).

Our computerised diagnostic-dynamic procedures include a novel implementation of L2 English reading tasks focusing on specific reading constructs (based on prior research and stakeholder surveys), where learners receive mediation based on DA whenever they encounter difficulties in answering the questions. The support becomes increasingly more explicit and detailed with each unsuccessful attempt until the correct answer is revealed to the learner. To complement the mediation, a chatbot is made available to the learner as soon as they answer the question correctly (after the first or subsequent attempts) so that they can ask the bot about the text or the task.

In addition to the chatbot, AI is used, for example, to vary the phrasing of mediation messages to make the stardardized mediation less repetitive. To support the students' reading development, the system also creates automatic exercises (with accompanying mediation) from texts uploaded to it by the teacher or student for practicing the more linguistically oriented grammatical and lexical constructs that contribute to reading comprehension (e.g., discourse markers, tenses).

We report on findings from using the online reading tasks with over 300 learners in 2025 focusing on learners' interaction with the system and their, and their teachers', feedback on the relevance and usefulness of the tasks, mediation, and chatbot.

Paper 8: Examining AI Chatbots' Efficacy in Formative Writing Assessment for EFL Students' Skill Enhancement

Aaisha Al Balushi, University of Technology and Applied Sciences

This study examines the effectiveness of AI chatbots as formative assessment tools in improving the English writing skills of foundation program students at a higher education institution in Oman. With the increasing integration of technology in language education, AI-powered tools like ChatGPT, DeepSeek, and Grammarly offer promising opportunities to support Omani students who often struggle with grammatical accuracy, coherence, and academic writing conventions.

Using a mixed-methods design, this research assesses the impact of AI chatbot feedback on students' writing development over a 12-week foundation course. Quantitative analysis of pre- and post-intervention writing samples measures improvements in linguistic accuracy, structural organization, and vocabulary use. Qualitative data from student surveys and instructor interviews explore perceptions of AI tools, including their accessibility, relevance to academic writing tasks, and effectiveness in addressing common challenges faced by Omani learners, such as L1 interference and limited exposure to English.

Findings indicate that AI chatbots enhance student engagement by providing instant, personalized feedback, enabling learners to identify and correct errors independently. However, the study also reveals limitations, including chatbots' occasional inability to grasp culturally specific expressions and the risk of students becoming overly dependent on automated corrections. The paper concludes with practical recommendations for optimizing AI chatbot use in Omani foundation programs, such as integrating them as supplementary tools alongside instructor feedback and designing targeted training to help students critically evaluate AI-generated suggestions.

Paper 9: Evaluating Chatbot Authenticity in Simulations of Spoken Interaction: Demonstrating the Utility of Corpus-based Methods for Development and Validation

Dana Gablasova, Luke Harding, Vaclav Brezina, Emil T. Hazelhurst, Lancaster University;
Barry O'Sullivan, Richard Spiby, British Council

Spoken dialogue systems, including AI-powered chatbots, have the potential to revolutionise the teaching and assessment of second language speaking. However, the use of chatbots powered by Large Language Models (LLMs) such as ChatGPT raises a crucial question: how authentic is the language produced by the chatbot interlocutor? Inauthentic language use in chatbot productions would risk weakening the validity of speaking tasks used for assessment or pedagogical purposes (Voss & Waring, 2025; Xi, 2023). Methods are therefore required to generate validity evidence to evaluate claims concerning the authenticity of chatbot productions.

In talk, we present a methodological framework for applying corpus linguistics to systematically evaluate the authenticity of chatbot production in relation to (spoken) production in a general target language use domain. We demonstrate the approach through data drawn from an illustrative case study: a low-stakes formative assessment system in which learners interact with a ChatGPT-powered bot. A Chatbot Corpus containing approx. 290,000 words from 600 simulations representing target ChatGPT production was created, representing two GPT versions (3.5 and 4), and three "temperature" settings. This corpus was then compared with relevant sub-corpora in the British National Corpus 2014, which contains 100 million words of British English collected in naturalistic settings. Analyses were conducted at macro- (multidimensional analysis), meso- (comparative frequency analysis), and micro-levels (occurrence of specific pragmatic feature analysis). Findings demonstrated that ChatGPT-powered chatbot production was systematically more similar to genres of written rather than spoken communication: output demonstrated higher lexical density and was characterised by a relatively low occurrence of features typical of spoken

communication such as stance and pragmatic markers. We argue that the methodological framework is applicable across different chatbot models, allowing researchers and developers to use this approach with newer, more refined AI-powered conversational agents in the future.

Paper 10: Effects of Script Presentation Mode on Raters' Scoring Outcomes and Cognitive Processes: An Eye-tracking Study

Daniel Yu-Sheng Chang, University of Bristol

This eye-tracking study examined how script presentation mode affects raters' scoring outcomes and cognitive processes in the two on-screen marking conditions: scanned handwritten and online word-processed scripts. Fifty-three raters scored 30 scripts, once in word-processed form and once in handwritten, across four criteria. During rating, their eye-movement data and scoring order/adjustments were concomitantly recorded. Other sources of data included raters' mental effort, use of the rating scale, self-reported scoring processes and perceptions across modes. Data were analysed using any-Facet Rasch Measurement, multilevel modelling, and the inductive qualitative approach.

Results of score analyses showed that 80% of raters demonstrated satisfactory rater fit. Rater severity spanned 3.09 logits, with grammar scored most severely and organisation most leniently. For the same script across mode, word-processed scripts were statistically scored lower than their handwritten counterparts for both the overall score and the four analytic sub-scores. Considering variability for script and rater respectively, mode effects seemed not to be systematic. While no certain word-processed scripts were severely marked for their mode, individual raters varied in how they mark severely or leniently word-processed scripts.

In terms of cognitive processes, both pupil size and self-reported mental effort indicated that raters invested significantly higher cognitive load in handwritten scripts. Raters' eye-movement patterns seemed to differ across modes. Specifically, word-processed scripts prompted more visits, likely due to clearer layout and easier navigation. Raters generally assigned scores following the sequence of the rating scale. However, when salient features (e.g., noticeable errors) relevant to specific criteria occurred, raters deviated from this sequence towards scoring the criteria associated with those features. The frequency of score adjustments appeared to be associated with raters' severity.

This study would contribute to fuller understandings of script presentation mode in L2 writing assessment and offers insights for rater training, rating condition and GenAI-powered scoring development.

Paper 11: Understanding Cognitive Processes Underlying Multiple-text Items Through Eye-tracking and Verbal Protocol Methods

Aylin Unaldi, University of St Andrews

As our understanding of language use continues to broaden, language assessment seeks to reflect these conceptual developments. In recent years, there is increased interest to introduce multimodal and multiple-text components to align more closely with the cognitive demands of real-world comprehension. One notable response to this shift is the incorporation of complex reading tasks that require the integration, comparison, evaluation, and synthesis of information across several sources. Such tasks are particularly common in academic contexts (Moore et al., 2010) and are associated with deeper learning and more advanced reasoning (Cerdan, 2006; Gil et al., 2010).

Several English for Academic Purposes (EAP) assessments now claim to operationalise multiple-text reading skills. However, there is currently no published evidence examining the degree to which these tests successfully capture the construct. This study investigates how multiple-text reading is

represented in selected English proficiency exams, including ISE II, MET, ECCE, and SAT. It explores whether the subskills and strategies outlined in these tests correspond to theoretical accounts of documents model reading, and whether the tasks actually elicit the kinds of processing associated with this model.

Data were collected from 12 participants through eye-tracking and retrospective think-aloud protocols. Preliminary analysis suggests that while some test tasks appear to engage genuine documents model strategies, others tend to prompt more basic, search-oriented behaviours. Based on these findings, a preliminary taxonomy of multiple-text reading strategies was developed, grounded in theoretical models of reading across texts. It aims to guide future test development for more cognitively grounded assessment of multiple-text comprehension.

Paper 12: Teacher Literacy in IELTS Writing Assessment: Applying Task 2 Band Descriptors in Formative Feedback

Yasaman Ibayeva, University of Bristol

In high-stakes English language testing contexts such as the International English Language Testing System (IELTS), teachers play a central role in preparing learners to meet assessment expectations. However, little is known about how well teachers understand and apply official IELTS assessment criteria when giving formative feedback. This study explores IELTS writing teachers' feedback literacy, focusing on their ability to interpret and use the Task 2 band descriptors to guide student improvement.

Adopting a sociocultural view of teacher assessment literacy, the study investigates how teachers engage both conceptually and practically with the four Task 2 criteria. The research involved sixteen IELTS teachers - ten based in Azerbaijan and six in the UK - who provided written feedback on student essays. This generated a corpus of 121 compositions for quantitative analysis. Additionally, semi-structured interviews offered qualitative insights into teachers' interpretations, feedback beliefs, and contextual factors shaping their practice.

Teacher comments are analysed across dimensions, including feedback purpose, approach, focus, tone, and alignment with IELTS criteria. The study also considers teachers' stated intentions and interpretations to better understand how formative feedback is constructed in IELTS classrooms.

Preliminary analysis reveals emerging patterns in how teachers apply the band descriptors, which will be elaborated upon during the presentation. This research contributes to work on teacher feedback literacy and offers implications for training, development, and assessment-informed feedback. By examining teachers' use of IELTS criteria in two contexts, the study highlights equity and consistency in formative feedback for high-stakes assessment.

Paper 13: Online EMI Interaction in University Classes: Developing Evidence-Based Resources

Fumiyo Nakatsuhara, Chihiro Inoue, Katherine Halley, CRELLA, University of Bedfordshire; Akiko Kiyota, Tokyo University of Foreign Studies; Ridhwan Abdullah, Univerisity of Reading Malaysia; Hasif Jazila, Amilin Nordin, Multimedia University, Yasuyo Sawaki, Waseda University

Regardless of whether language assessment is used for gatekeeping purposes at the entry stage of English Medium Instruction (EMI), its value is increasingly recognised for learning-oriented purposes and for raising policymakers' and stakeholders' awareness of the importance of language proficiency (Galloway, 2020; Hultgren et al., 2022). While the interwoven relationship between content and language in EMI, student characteristics, and teachers' language assessment literacy all contribute to the complexity of developing and implementing language assessment for EMI programmes, what remains essential as a first step is to deepen our understanding of the target language use domain,

to be reflected in test specifications.

The research reported in this presentation is part of a large-scale interdisciplinary project that investigated the nature of interaction in 24 video-recorded online EMI classes at two universities in Malaysia and Japan. Multimodal transcriptions were developed to capture screen activity, teachers' and students' spoken contributions, and chat interactions, allowing for an understanding of their simultaneous and sequential relationships.

Using an extended version of O'Sullivan et al.'s (2002) function list—including 12 informational, 19 interactional, 3 interaction management, and 3 technology-use functions—we identified frequently occurring language functions across different communication modes and compared findings between the two EMI contexts. Discourse analysis based on the multimodal transcriptions illustrated how multiple functions were combined and realised in different ways for specific communicative purposes.

Informed by these findings, we designed prototypes of low-stakes online EMI readiness test tasks that reflect insights into the nature of online EMI discourse gained from this study. In addition to exemplifying a model for transforming empirical findings into practical resources, this research aims to help teachers develop and tailor test tasks, empowering educators and fostering collaboration between subject specialists, EAP teachers, and language testing researchers.

Paper 14: Documenting Validity: A Review of Test Manuals for High-Stakes English Language Tests

Ben Naismith, Ramsey Cardwell, Masha Kostromitina, Duolingo

Test manuals are central to how test developers communicate key assessment features and details to a wide range of stakeholders, including researchers and educators. For policy makers, test manuals are particularly important in evaluating language tests for use under relevant policies. These documents also serve as a primary mechanism for demonstrating test validity, broadly defined as the extent to which evidence and theory support the intended interpretation and use of test scores.

However, while considerable effort goes into producing test manuals, there is no consensus on what exactly they should contain, despite guidance from assessment experts and organizations. In this presentation we explore this issue through a comparative review of publicly available test manuals (or equivalent documentation) for ten high-stakes English language proficiency (ELP) tests. In doing so, we aim to better understand cross-organizational similarities and differences in test manual content and structure, to reveal trends in how ELP test developers operationalize validity documentation, and to suggest improvements to existing guidelines for test manuals.

Using the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014) and elements of Bachman and Palmer's (1996) test usefulness framework, we developed a coding scheme to identify how validity-related information is presented across test facets. Our analysis revealed substantial overlap in the types of validity evidence included, with many manuals addressing core areas such as test purpose, construct definitions, and scoring procedures.

However, variation was also observed in the depth of reporting across different dimensions, and certain important aspects were less consistently addressed, especially consequences of test use and fairness considerations. In addition, many ELP tests do not have specific, downloadable, test manuals, instead opting to put all test information on their websites in multiple locations. We discuss implications for transparency, comparability, and evolving best practices in documentation for modern high-stakes ELP assessments.

Paper 15: Assessing Metapragmatic Awareness for Intercultural Communication: Harnessing Interdisciplinary Power in Language Assessment

Shishi Zhang, University College London

Despite the past five decades' efforts in assessing second language pragmatic competence, challenges remain, including assessing skills more transferrable mirroring test takers' real-life language needs, and designing a purpose-built assessment to operationalise such a construct. Addressing this research gap, this study proposes the assessment of metapragmatic awareness, namely, test takers' ability to calibrate utterances and meanings reflexively and dynamically, considering the interlocutor, communicative needs and interactive goals while mediating linguistic/cultural differences to establish a common ground for communication (Kecskes, 2014, 2022; McConarchy, 2018). Within the field of applied linguistics, there are diverse interpretations of metapragmatic awareness, albeit using the same "metapragmatic" label. A few studies trying to assess metapragmatic awareness are actually assessing elements more in the realm of pragmatic awareness without emphasising reflectivity and reflexivity of test takers during communication. Therefore, this study is the first to assess "meta"pragmatic awareness. Specifically, the study contextualises the assessment to UK pre-sessional students.

Drawing on interdisciplinary approaches from fields such as SLA and psychology, the study designed an innovative task format: video analytical tasks, which were finalised through a multi-stage exploratory sequential mixed methods design (Creswell & Plano Clark, 2018). This presentation explains the process of developing and initially validating the tasks by systematically engaging key stakeholders (e.g., pre-sessional students and teachers), following the Socio-Cognitive Framework (O'Sullivan & Weir, 2011). Task-specific analytic rating scales were developed, drawing on both test-external resources (language proficiency frameworks, expert intuitions, existing scales) and test-internal resources (rater performance; Knoch et al., 2021). Many-facet Rasch analysis of field testing data suggested that the tasks and rating scales were functioning well for the target test population. Implications will be discussed on encouraging systematic pragmatics instruction in UK pre-sessional programmes, and harnessing the power of metapragmatic awareness in fostering intercultural communicators to better handle communication on increasingly globalised UK campuses.

Paper 16: Applying the Lessons from Second Language Test Validity Frameworks to Tests of English as a First Language

Nicola Latimer, Fumiyo Nakatsuhara, Sathena Chan, Pauline Madella, CRELLA, University of Bedfordshire; Samantha Orciel, Lydia Ridding, OCR, Cambridge University Press and Assessment

This presentation responds to the theme of equ(al)ity, diversity, and inclusion (EDI) in language assessment and testing. In the UK, GCSEs are designed to be taken by 15 and 16-year-olds and are intended to open the door to career and further educational opportunities. However, results consistently reveal inequalities according to geographic location, the type of school attended (DfE, 2024), and ethnicity (The Health Foundation). Those who leave school without these basic qualifications often struggle to access further educational opportunities, including vocational courses or job opportunities (British Dyslexia Association, 2023).

Some exam boards are planning reforms to the way some subjects, including GCSE English Language, are assessed. When such changes are made, it is critical to consider the knowledge and processes being tested and ensure they align with those the test intends to target. It is also important to consider how individuals from diverse cultural and ethnic backgrounds, as well as those with special educational needs, will fare on the test.

This presentation presents the findings of a joint baseline research project by a team of language testing researchers and a GCSE English Language Exam Board to analyse the context and cognitive

validity of the reading section of a GCSE English Language exam using a set of predefined parameters. The presentation explains how a framework, originally developed for reviewing tests of English as a second language, was adapted and developed for application in an L1 context. The presentation will focus on how textual analysis techniques and expert panel reviews were used to evaluate the context and cognitive validity of the exam.

The findings of the project are reported along with the implications for future versions of the L1 English exam. We will also discuss implications for better EDI practices in L1 and L2 assessment contexts.

Paper 17: Optimizing Computer Adaptive Test Design Through Simulation: Investigating the Oxford Placement Test

Nathaniel Owen, Oxford University Press

This research presents a novel adaptive test simulator addressing computer adaptive testing optimization through computational simulation methods. The study investigates the comparative performance of different test configurations, and utility of grouped tasks (testlets) within the adaptive Oxford Placement Test. The simulator was developed in Python v3.10 and implements the Rasch model, ability estimation convergence and standard error progression. Using live test data from 10,361 test takers, three test design configurations were evaluated against the original dataset and baseline 42-item configuration, comparing ability estimates and standard errors across the ability range using metrics including correlations, Cohen's κ, and RMSE. Key findings demonstrate that careful test length reduction maintains psychometric integrity while improving administrative efficiency. A 32-item configuration emerged as optimal, showing strong measurement characteristics (SEM ≤ 0.4 logits) and robust classification accuracy (84.7% exact agreement at half-CEFR bands, κ = 0.83), reducing test length by 24%. Removing testlets in favour of discrete items revealed minimal overall impact on measurement precision, though increased error at higher proficiency levels suggests testlet tasks discriminate effectively among high-ability candidates (Chalhoub-Deville, 2001; Wainer et al., 2007).

This research bridges computational psychometrics and language testing, demonstrating how simulation-driven optimization can transform adaptive test design. The findings establish that data-driven approaches can achieve substantial efficiency gains while preserving measurement integrity, offering a replicable framework for evidence-based test development across diverse language assessment contexts.

Paper 18: Human-AI Collaboration in Language Testing: Equity and Validity

Ramsey Cardwell, Alina von Davier, Jill Burstein, Duolingo

The integration of generative artificial intelligence (AI) into language assessment presents both significant opportunities (e.g., scaling test content production) and ethical challenges (e.g., ensuring fairness to all test takers). Human-in-the-loop (HiTL) AI, a collaborative approach integrating human expertise with AI-driven technologies, is increasingly recognized for safeguarding validity and fairness in language testing while leveraging AI's affordances. Using the Duolingo English Test (DET) as an illustrative example, this talk introduces an innovative ecosystem-based approach known as the "item factory", which integrates human expertise and AI capabilities, facilitating scalable, efficient, and high-quality test development processes. Drawing on principles of intelligent automation from engineering and manufacturing, this approach effectively combines evidence-centered test design, natural language processing (NLP), and computational psychometrics to support rigorous validation and quality assurance.

The presentation will also discuss Human-Centered AI, a broader framework that advocates for the

inclusion of educators, assessment experts, and students in the test development process, thereby ensuring AI systems augment rather than replace human roles. This framework highlights transparency, fairness, ethical considerations, and the role of trust in adopting AI-driven assessments. Additionally, the DET's Access Program is presented as a practical example of how ethically aligned HiTL-AI frameworks can support greater inclusivity and global accessibility in language testing. The presentation concludes by exploring future research directions and practical implications for integrating interdisciplinary, inclusive, and ethically sound HiTL-AI practices into high-stakes language assessments.

Paper 19: Targeting the Top: Evaluating Differentiation Among High-performing Candidates in Items Designed to Stretch and Challenge

Ana Ulicheva, Sarah R. Hughes, Pearson

This study investigates the effectiveness of two innovative item types, Respond to a Situation and Summarise Group Discussion, in differentiating high-performing candidates in a high-stakes language proficiency test. These tasks were designed to elicit more cognitively demanding and linguistically complex responses, targeting higher-order language skills such as discourse management, pragmatic flexibility, interactional competence, and synthesis.

Using secondary analysis of field test data from 2,690 adult test-takers, we explored whether these research item types enhance score differentiation beyond what is captured by standard speaking tasks. Test-takers completed both standard items and the two new tasks, with responses scored analytically for content, fluency, grammar, and vocabulary. We compared two matched groups: Group A, high scorers on the research items, and Group B, individuals with comparable performance on standard tasks but lower scores on the research items.

Our analysis followed four stages: (1) descriptive and entropy-based evaluation of score distributions with and without research items, (2) classical test theory and psychometric analysis of item characteristics, (3) comparison of score profiles across traits emphasizing advanced skills, and (4) linguistic analysis of spoken responses using metrics such as lexical diversity, syntactic complexity, and cohesion markers.

Findings indicate that the research item types successfully increase score entropy and item differentiation, improving the test's capacity to distinguish between high-proficiency individuals. Group A's responses demonstrated greater linguistic complexity and more sophisticated discourse strategies. These results support the construct validity of the new tasks and their role in eliciting richer, more authentic language use. The study underscores the value of integrating cognitively demanding items in high-stakes assessments to better capture advanced proficiency. It also highlights the benefits of analytic scoring and linguistic analysis in enhancing performance interpretation. The findings inform both test development and validation practices in language assessment.

Paper 20: Beyond the Lab: A Virtual Desktop Data Collection Method for Writing Process Research

Michelle Czajkowski, Radboud University

Academic writing by students typically occurs in familiar, resource-rich environments. Yet assessments of academic writing ability often take place under unfamiliar, timed, and resource-restricted conditions. This mismatch raises concerns about ecological validity, construct representation, and assessment fairness (East & Slomp, 2023; Pusey & Butler, 2024). An important step in addressing these concerns is to understand academic writing processes in situ.

This presentation introduces a remote data collection procedure designed to capture writing process data in students' natural study environments. First-year undergraduate students were granted virtual access to a lab computer equipped with keystroke logging and screen capture software. This setup enabled unobtrusive collection of process data, including students' use of online resources, without requiring software installation on personal devices—effectively extending the experimental environment for writing process research (Wengelin et al., 2019).

A somewhat novel data collection solution (Denvir, 2017), the virtual desktop procedure used in this study was developed through consultation with data managers, IT security and privacy officers, and research ethics personnel, resulting in a manual refined through participant and researcher feedback. Amid growing interest in and concern over remote data collection in social sciences and health research (Roberts et al., 2025; Lobe et al., 2022), and increasing attention in language assessment to writing process research, this method offers a practical model for future studies, supporting the field's understanding and assessment of academic writing.

In this talk, I outline the method and the challenges encountered in its implementation. I also present observations from student feedback and researcher notes. The discussion will be framed within ongoing debates about ecological validity in writing research and assessment, with implications for both assessment design and the construct of academic writing.

This presentation will be of interest to researchers and practitioners seeking a practical, methodologically grounded approach to collect writing process data remotely.

Paper 21: Can Grammatical Metaphor be Used as a Marker of Writing Complexity in L2 Assessment?

Nicholas Glasson, Andrew Kitney, Cambridge University Press & Assessment

The concept of grammatical metaphor (GM) (Halliday, 1985)—where actions or qualities are re-construed as nominalised "things"—offers a powerful lens for examining meaning-making in academic writing. Like other metaphors, GM allows writers to "represent something as something else" (McGrath & Liardét, 2023, p. 33). Despite its relevance to meaning-based writing complexity, GM is often overlooked in favour of surface-level measures such as lexical diversity and syntactic range (Yasuda, 2024). This study addresses that gap by illustrating how writing complexity can be explored through corpus-informed mixed-methods analysis using MAXQDA.

This session presents findings from 152 Linguaskill writing scripts across CEFR levels (below B1 to C1+), focusing on the frequency and use of GMs. Scripts were tagged semi-automatically using a pre-established GM inventory (McGrath & Liardét, 2023), enabling both quantitative and qualitative insights. The analysis examined correlations between GM usage and proficiency scores, and qualitatively explored, through a systemic functional grammar lens, how GMs were deployed across levels.

Findings show a moderate positive correlation between GM frequency and proficiency, with process-to-thing shifts (e.g., transform → transformation) most common. GM usage increased with proficiency, suggesting a developmental trajectory. These results highlight GM's potential as an indicator of academic discourse competence in L2 writing, with implications for instructional design and assessment - specifically in rebalancing form-based and meaning-based complexity in construct definitions.

Paper 22: AI-generated Voices and Accent Diversity: Towards Fairer Listening Assessment

Leyla Karatay, Nicholas Glasson, Andrew Mullooly, Cambridge University Press & Assessment

The growing diversity of accents encountered in English-speaking contexts highlights the need for L2 listening assessments to reflect this linguistic reality (Ockey & French, 2014).Given the logistical challenges of sourcing authentic accents for assessment tasks, AI-generated accented voices offer a promising alternative.Drawing on the concept of Global Englishes (Rose et al., 2009), this study uses AI-generated accents from different English varieties to mirror the linguistic diversity.As the first phase of a larger project, this study evaluates the effectiveness of a text-to-speech tool in producing different accented English varieties for a monologue-based listening note completion task type.Three human actors (French, Indian, and British accented) recorded scripts for the same three tasks., which were then replicated using AI generated voices selected for accent similarity.

Twenty experts from diverse language backgrounds evaluated all recordings for intelligibility, comprehensibility, and accent strength. For intelligibility, they transcribed 46 segments from 20-second excerpts of three AI-generated and three human-recorded audios using the same scripts (Munro&Derwing,2020).Comprehensibility and accent strength were rated on a 9-point scale (Munro & Derwing, 1994),and participants indicated whether each voice was AI- or human-generated. Transcriptions were coded using an adapted Munro and Derwing (2020) scheme, covering exact match, substitution, novel words, and omissions. Two raters independently coded the data to examine transcription accuracy and error patterns across voice types. For comprehensibility and accent strength ratings, descriptive analysis was used to explore how the AI-generated and human-recorded voices were perceived. Results showed AI-generated voices were perceived as highly similar to human ones, with many experts unable to reliably distinguish between them-suggesting potential in using AI-generated voices to create diverse audio at scale.

Paper 23: At the Intersection of Sociolinguistics and Language Testing: Comprehensibility, Pronunciation Features and English as a Lingua Franca

Sheryl Cooke, British Council, also affiliated with University of Jyväskylä; Karen Dunn, British Council

This study addresses the role of comprehensibility in a language testing setting, and, more specifically, the interdisciplinary tension arising between norm-informed language assessment and the fluidity of sociolinguistics in the context of English as a Lingua Franca (ELF). Despite the move from outdated 'native speaker' (EL1) models towards the goal of communicative effectiveness, many English language tests continue to use EL1 norms, entrenching inequity and perpetuating unfairness.

ELF is characterised by high degrees of variability, with pronunciation being the most obvious difference for the listener. Numerous studies have attempted to extricate the phonological features contributing to comprehensibility in English (e.g., Kang, et al., 2020) largely drawing on judgements from EL1 speakers and/or expert raters. This study, meanwhile, investigates features associated with both EL1 and EL2 novice listeners' comprehensibility ratings. Since most ELF interactions do not include EL1 speakers, this better reflects the communicative reality.

56 speech samples from Chinese L1 speakers of English were rated for comprehensibility by 30 listeners: 15 EL1 and 15 EL2 speakers with high levels of English proficiency. Speech samples were analysed across five categories of phonological features: segmental accuracy, syllable accuracy, word stress accuracy, vowel reduction ratio, and functional load substitutions. These linguistic features were tested for their relationship with listener ratings within a generalised linear mixed model (GLMM) framework. Whilst the CEFR levels allocated by the rating scales were found to account for a large portion of variability in ratings, findings additionally show that EL2 listeners were likely to rate more positively for comprehensibility than their EL1 counterparts. Of the linguistic features, only segmental error played a significant role in the model.

This empirical study highlights the imperative for language assessment researchers and practitioners to attend to the sociolinguistic realities of language use and makes recommendations for teaching and testing.

Key words: comprehensibility, ELF, phonological features, linguistic equity, GLMM

Paper 24: Test Anxiety Revisited: A Comparative Study of At-home vs In-centre High Stakes Tests

Ruolin Hu, University College London

Language testing, especially when tied to high-stakes outcomes, often prompts considerable anxiety. While test anxiety has been extensively researched, studies have generally focused on traditional linear, in-person assessments administered at test centres or in schools. The growing adoption of adaptive, at-home test delivery raises important questions about how this development may mitigate or reshape test anxiety.

This study investigates test anxiety in an at-home adaptive test compared to two in-centre tests. Using a 28-item retrospective self-report questionnaire, we collected data from 492 test takers who completed both the at-home and one of the in-centre tests. The questionnaire measured two primary dimensions of anxiety—worry and test-irrelevant thinking—across three key stages: pre-test, during-test, and post-test.

Preliminary findings reveal that participants consistently reported experiencing lower anxiety levels when taking the at-home test, with significantly reduced worry and fewer test-irrelevant thoughts. Wilcoxon signed-rank tests confirmed significant differences across all comparisons between in-centre and at-home testing, with large effect sizes averaging .56 for test 1 and .79 for test 2 (in-centre vs at-home), indicating substantially higher anxiety during in-centre testing and results robust to Bonferroni correction.

To deepen this analysis, we link test anxiety to test performance data for all three tests. We found that while the overall effect of anxiety on test performance is similar across delivery modes, differences are revealed when test stages (i.e. pre, during and post) are considered. We build predictive models to further unpack test anxiety by considering test taker characteristics such as English level, age, gender, race, and their first language.

Findings are discussed in relation to the construct validity of the at-home adaptive test and the broader implications for fairness and access in language testing, particularly as remote testing becomes increasingly central to high-stakes decisions in global admissions and immigration contexts.

Paper 25: Phraseological Knowledge in a Dialogic Speaking Task Across Assessed Levels of Proficiency

Parvaneh Tavakoli, University of Reading; Takumi Uchihara, Tohoku University; Svetlana Mazhurnaya, Phil Smyth, University of Reading

Despite its significant role in successful communication (Kremmel et al., 2017), phraseological knowledge is not assessed in most tests of speaking, and its relationship to proficiency remains under researched. Previous research (Eguchi & Kyle, 2020) has shown that using a greater number of high-frequency and sophisticated multiword sequences (MWSs) is positively, and largely linearly, linked to proficiency in monologic task performance. To date there is no evidence to suggest use of MWSs is similarly linked to proficiency level in dialogic performance. To help fill this gap, the current study aimed at examining the use of MWSs in a dialogic task across assessed levels of proficiency in

Test of English for Educational Purposes (TEEP).

The data set comprises 127 speech samples from speakers at six levels of the TEEP from 5.0 to 7.5 (equivalent to A2-C1 CEFR levels) performing a dialogic discussion task. The audio data were transcribed and subjected to MWSs analysis in TAALES (Kyle et al., 2018) for bigram and trigram measures of frequency, proportion and association strength. In addition, a phraseological knowledge scale (accuracy, diversity, sophistication and abundance of MWSs) was developed against which each audio recording was rated by three experienced raters.

The results of a series of ANOVAs, used to investigate the relationship between proficiency and MWSs, suggested that none of the n-gram measures were associated with levels of speaking MWSs, suggested that none of the n-gram measures were associated with levels of speaking proficiency. The results of a linear mixed-effects modeling, used to analyse the pair-level use of MWSs, suggested that pair-level effects explained 40% and 30% of the total variance for bigram and trigram proportion, respectively. Further analyses revealed that lower-proficiency speakers in a nationality-matched pair tended to produce a greater proportion of bigrams and trigrams than those in a nationality-unmatched pair. These findings challenge the central role of MWSs documented in monologic task performance and suggest a complex interplay of phraseological competence and speaking performance when dialogic processing is considered. Raters' judgement of test-takers' phraseological knowledge, however, demonstrated high positive correlations (minimum $r = .747$, $p < .001$) with assessed levels of proficiency, suggesting human raters' judgement is linked with phraseological knowledge.

Poster 1: Investigating Cognitive Process in Duolingo English Test Listening Section: An Eye Tracking Approach

Jia Li, Yuxuan Yang, University of Southampton

The landscape of university entry requirements and English proficiency testing has changed significantly, particularly with the increased demand for remote education and assessment since 2020. Many higher education institutions worldwide now accept the Duolingo English Test (DET) as a valid measure of English proficiency. However, there remains a lack of empirical research focused specifically on the DET.

This study employs eye-tracking technology to explore candidates' cognitive processes during second language listening comprehension within the computer-based, adaptive DET. In this format, the test dynamically adjusts the difficulty of items to more effectively estimate a test taker's language ability.

Our investigation focuses on Part One and Part Two of the Interactive Listening section. In Part One, test takers listen to a short scenario and type the missing word according to the questions; in Part Two, they complete a longer dialogue by selecting the most appropriate option. We examine how test takers allocate visual attention and how their attentional patterns relate to test performance and language proficiency levels. By analysing eye-tracking data—including first fixation time, fixation duration, saccadic paths, and pupil dilation—along with post-test interviews, we aim to gain insights into test-takers' listening strategies.

The objectives of this study are to:
(1) Identify the cognitive load associated with computer-based listening test items; and
(2) Investigate the validity of test-taker performance within a digital test environment.

The findings will contribute to the development of more effective computer-based language assessments and inform best practices in second language teaching and learning.

Poster 2: Comparing Overall Score Distributions in the Evaluation of an Automated Scoring System

Edmund Jones, Trevor Breakspear, Jing Xu, Cambridge University Press & Assessment

Automated scoring is widely used for constructed responses in language assessment, and many researchers are now interested in using generative AI for this purpose. The accuracy of an automated scoring system is usually investigated by measuring the agreement between scores from the automated system and reference scores from reliable human examiners. However, this may not be adequate to meet regulatory requirements for high-stakes tests.
One set of guidelines states that automated systems should additionally produce similar overall score distributions to human examiners. For this comparison visual judgement is essential, and it is also useful to have an objective measure of distribution similarity. The question of what measure to use has not been thoroughly debated. Researchers sometimes report the mean automated score and mean reference score. These are easy to understand but not sufficient as two probability distributions can have the same mean but different shapes.

We will introduce the use of earth-mover's distance (EMD) for comparing two score distributions. The intuitive idea is to regard the two distributions (like the bell curve for the normal distribution) as two piles of earth; EMD is the minimum "cost" of turning one pile into the other, where the cost of moving a small piece of earth is the amount of earth times the horizontal distance by which it is moved.

EMD was introduced in the field of computer vision and has been used in biology, geology, and machine learning. It has properties that make it suitable for comparing test score distributions and it is interpretable to some extent. We will show properties of EMD, illustrate it with several datasets including one from a high-stakes test of English writing proficiency, and compare it with other statistics.

Poster 3: "Intelligent Enough to Assess Me?": Exploring the Intelligence and Effectiveness of AI Feedback in Speaking Language Assessment

Yuchen Xing, University of Southampton

As generative AI (GenAI) becomes increasingly integral to language learning, the quality of its automated feedback is a key concern. Feedback from large language models (LLMs), when guided by simple prompts, often remains generic, failing to address a learner's individual needs or proficiency level. This study moves beyond simple prompts to investigate how strategic prompt engineering can improve the pedagogical value of AI-generated revisions for second language (L2) speaking.

This research is based on a corpus of secondary data, comprising 15 authentic IELTS Speaking Part 2 (long turn) transcripts. In the first phase of the study, three distinct GenAI tools, guided by three different prompts, were used to generate revised, higher-scoring versions of each original text, producing a total of 135 revised texts. Subsequently, a separate GenAI tool was then used to quantitatively measure the effectiveness of these 135 revisions by calculating a 'score improvement'. Following the data collection and analysis, a follow-up analysis will be conducted on a selection of texts exhibiting notable characteristics within the data to investigate their key linguistic and pedagogical features. Thereby, providing an explanation for the quantitative findings of the initial phase.

This study will address two core questions: (1) Which AI models and prompts produce the most effective revisions of L2 speaking texts? (2) What linguistic and pedagogical features characterize the most successful AI revisions? Through a close analysis of prompt engineering and AI-generated revisions, this research offers an objective framework for evaluating GenAI feedback, provides

practical guidance for learners and educators engaging with AI technologies, and contributes to an evidence-based understanding of automated language assessment.

Poster 4: Speaking On-topic: An Enhancement to the Socio-cognitive Framework and Raters' Perceptions about Feelings

Judith Fairbairn, University College London

This poster explores findings from a mixed methods PhD study on (1) the internal criteria that 30 raters report using to decide if a speaking response is on- or off-topic and (2) the relationship between these internal criteria and raters' Aptis scoring and on-topic decisions with respect to rating severity, reliability, consistency and construct relevance.

A key recommendation discussed in this poster is to add topic relevance decisions by raters to the socio-cognitive framework (Weir, 2005). The framework includes rater reliability in general but does not specifically discuss how raters deal with topic relevance decisions made by test-takers. The focus of the framework is also mostly on spoken interaction with a human interlocutor. There is scope from this research to enhance the socio-cognitive framework by including rater perceptions of topic relevant decisions made by test-takers in a semi-direct speaking test.

This poster also elaborates on one finding that some raters may struggle with topic relevance decisions for spoken responses about feelings. The CEFR Companion Manual (Council of Europe, 2018) Sustained Monologue: Describing Experience scale has a high B1 level descriptor "Can clearly express feelings about something experienced and give reasons to explain those feelings" (p. 70). Raters may not fully understand how to measure topic relevance for prompts that ask test-takers to describe their feelings.

Poster 5: PFL Proficiency Assessment: Factors Influencing Learner Performance in Different Cloze Test Formats

Clara Setas, Universidade do Minho (CEHUM)

This study explores the use of different variants of the gap-filling cloze format in the assessment of language proficiency in Portuguese as a Foreign Language (PFL), addressing a gap in empirical studies on varied assessment methods across age and context (Zhou & Li, 2022). Building on Bachman and Cohen's (1999) framework for language test validation, this study seeks to inform more efficient development practices by analyzing the correlation between learner performance and influencing factors, both individual (sociolinguistic profile, self-assessment, performance and strategic skills) and test-related (response format, item complexity and cognitive complexity). This study will employ three rational cloze test formats, in construct-equivalent versions, administered to PFL learners in formal learning contexts: open/structured response, multiple choice, and word association from a list, all based on narrative texts. Item selection is guided by linguistic structure type and complexity, defined according to Referencial Camões PLE (2017), language acquisition literature, and pretesting validation results using psychometric item-level analysis. The aim is to analyze (i) the validity and reliability of the different cloze test variants; (ii) the influence of response strategies on performance; (iii) the influence of sociolinguistic factors on performance; and (iv) the effect of cognitive load on performance. Data will be collected using questionnaires, surveys, think-aloud protocols and cloze tests in digital format. The study aims to contribute to informed assessment practices in PFL, particularly on how to shape and control item complexity in the development of cloze tests as a tool for measuring proficiency.

Poster 6: The Impact of Automated Writing Feedback on Self-Regulated Writing Strategies and L2 Writing Performance among Female EFL Learners at Saudi University

Maishael Alsanad, University of Southampton

As technology continues to shape language learning practices, the use of digital tools in L2 writing instruction has gained increasing attention. This study explores the impact of Automated Writing Feedback (AWF) on the development of self-regulated writing strategies (SRWSs) and writing performance among female EFL learners at a Saudi public university. In alignment with Saudi Arabia's Vision 2030 for digital innovation in education, the study evaluates the potential of Cambridge Write & Improve, an AI-powered feedback platform, to enhance both writing outcomes and learners' strategic engagement with writing tasks. A quasi-experimental design was adopted with 35 participants: 21 in the experimental group (receiving AWF via Cambridge Write & Improve) and 14 in the control group (receiving traditional teacher feedback).

Quantitative data were collected through pre- and post-writing tests and a validated SRWSs questionnaire that measures cognitive, metacognitive, motivational, and social strategies. The results revealed statistically significant improvements in the experimental group across all writing components, with the most notable gains in organisation and vocabulary. The control group also showed improvement in writing performance, but to a lesser extent. In terms of SRWSs, the experimental group demonstrated significant post-intervention gains in metacognitive planning and motivational strategies, whereas the control group's development in SRWSs was minimal and not statistically significant. Furthermore, the experimental group was found to outperform the control group in both writing scores and the use of SRWSs at the end of the treatment. These findings support the integration of AWF tools into EFL instruction as a means of enhancing both the product and process of L2 writing and provides practical implications for writing pedagogy in digitally evolving educational contexts.

Poster 7: Mexican and Brazilian High School EFL teachers' Perspectives: National English Policy and Its Impact on Classroom Assessment

Elsa Fernanda Gonzalez, Autonomous University of Tamaulipas; Gladys Quevedo-Camargo, University of Brasilia

In Mexico, the National English Program, launched in 1993 and fully implemented by 2005. At the high school level, the official curriculum aims for students to reach B1 proficiency after six semesters. Extensive research (Basurto-Santos & Weathers, 2016; Ramírez-Romero et al., 2012, 2014, 2016; Chablé, 2017; Sanchez-Menendez; Basurto-Santos, 2024) has revealed a persistent mismatch between national curriculum discourse and actual classroom practices. Teachers report struggles to meet these language policy goals due to structural challenges, lack of resources, and sociolinguistic diversity. In Brazil, the federal Base Nacional Comum Curricular (BNCC) (National Curricular Core - Brasil, 2018) similarly outlines objectives but often fails to consider regional inequalities, local languages, and the lived realities of students in public schools. The presentation describes the similarities and contrasts of Mexican and Brazilian national language policies and its impact on highschool EFL teachers´ assessment practices. Qualitative data from semi-structured interviews with Mexican and Brazilian English teachers in service in public high schools reflect on the tensions between language policy and classroom assessment practice, regarding socioeconomic backgrounds, access to technology, and the lack of culturally responsive materials. Thematic analysis of these interviews pinpoints the need to contextualize language assessment policies to the localized needs of teachers in each country so that classroom assessment is made an essential part of a Comprehensive Learning System (O'Sullivan, 2021). The poster presentation finalizes with a discussion of possible implications of the study and suggestions for future research.

Poster 8: Approaches to 'Testing the Untestable': A Year of Classroom-based Inquiry into Diverse Student Responses in Upper-secondary English as a Foreign Language (EFL) Literature Assessment

Lynn Williams Leppich, Bern University of Teacher Education

Practising teachers of upper-secondary English as a Foreign Language (EFL) in Switzerland are tasked with designing assessments which allow students to showcase not only their linguistic competence but also their understanding of literature and culture. While standardised frameworks designed to assess these necessarily offer a welcome attempt to introduce clarity, transparency and measurable outcomes, literature assessment also often invites – and arguably even demands – an openness to varied interpretations and modes of expressions.
A year on from my initial proposal to adopt a literature assessment mindset as a way of also assessing certain aspects of language, this poster presentation revisits and extends the discussion. Exploring what happened when these principles were systematically put into practice over the course of a school year, I will share how students responded to a diverse literary curriculum, with

tasks including group discussions, written responses, creative pieces and multimodal projects. In addition to showcasing student literary competence, these tasks also usefully provided multiple entry points for language use, encouraging learners to negotiate meaning, employ advanced vocabulary and construct coherent and nuanced arguments. At the same time, they challenged me as a teacher-assessor to maintain standards of rigour and fairness across a wider and more diverse spectrum of work.

My poster will share selected examples of student responses as well as reflecting on the principles and practices that supported meaningful assessment. Further, it will suggest practical strategies for scaffolding preparation, engaging students in co-constructing criteria and balancing flexibility with consistency in evaluation. Reflecting on the professional learning gained through this approach, ultimately I argue that literature assessment can meaningfully complement traditional language testing by modelling a potentially more inclusive and flexible conception of student achievement, one that honours voice, perspective, and critical and creative engagement.

Poster 9: A Journey into the Teaching and Assessment of Speaking in Portuguese Classrooms

Margarida Pato, CRELLA, University of Bedfordshire

This study examines an educational paradigm shift in Portugal and its impact on the way schools, and particularly teachers, conceptualise, implement and assess the curriculum, specifically in terms of speaking. It, thus, investigates the interplay between language policy, teaching, and assessment. By gathering context-specific data from three schools, the study also responds to Chalhoub-Deville and O'Sullivan's (2020) call for more context-embedded studies, while providing insights into teachers' assessment literacy and classroom-based speaking assessment, filling in a gap in research as identified by Fan and Yan (2020).

Anchored in O'Sullivan's Comprehensive Learning System (2020), which emphasises the interconnectedness of curriculum, delivery, and assessment, the research explores how these components interact in practice. Employing a mixed-methods approach, to achieve an in-depth understanding of a multidimensional reality, the study involved three English teachers, three Heads of Department, and three classes of B1 and B2 level secondary students in Portugal. Data were gathered through semi-structured interviews, classroom observations, teachers' weekly logs, students' formal assessment recordings, and coursebooks. The research conducted document analysis of curricular documents and guidelines, thematic analysis of interviews with Heads of Department and teachers, content analyses of direct classroom observations and assessment moments, and statistical analysis of teachers' logs and coded coursebook elements. Findings from

all data sources and analyses were triangulated to provide evidence of (dis)alignment of the three elements of the Comprehensive Learning System.
The findings aim to inform curriculum design policies, highlight the importance of the intersection between elements withing educational systems, underscore the influence of teacher beliefs on curriculum implementation, and elucidate the complexities of teaching and assessing spoken English. This research offers insights that may enhance teacher training and support more integrated approaches to language education, with implications also for countries where similar educational reforms are being implemented.

Poster 10: A Longitudinal Case Study Tracking Undergraduate EMI Learners' Potential Second Language Speaking Development in a Higher Education Context in Japan

Jamie Lesley, CRELLA, University of Bedfordshire

This longitudinal case study examines English Medium Instruction (EMI) in Japan's higher education (HE) context, where it targets undergraduate learners' potential second language (L2) development in one academic year of EMI studies. Today, EMI in HE is prevalent worldwide. Nonetheless, benefits for L2 development lack both evidence and effective assessments (Galloway & Rose, 2021), while L2 speaking is either deprioritized or absent in the few studies that exist. However, with increasingly multimodal communication and shifts in HE towards more integrated skills use (Khabbazbashi et al., 2023), providing coverage to all L2 skills with valid assessments helps evaluate EMI more thoroughly. Employing a mixed methods approach, this research explores L2 speaking at a Japanese university to determine the speaking skills required, the speaking gains, if any, students achieve in one academic year, and their documented learning experiences during this period. Accordingly, the study bridges identified gaps in the literature by exploring learners' possible L2 speaking development using a bespoke test informed by empirical data from the target context. The presentation details how syllabuses, rubrics, teacher surveys/interviews, and lesson observations established the speaking construct. Such findings informed the development of a bespoke test used 4 times with 12 students (with average speaking proficiencies of CEFR C1). To track (non-)progressions of L2 speaking, ongoing analysis comprises linguistic measures of fluency, complexity, accuracy, and total production, and qualitative analysis of interactional effectiveness. Students' reflections were collected through start-of-year surveys, followed by semi-structured pre-test interviews, and then end-of-year focus-group discussions. To the best of the researcher's knowledge, it is the first study to track speaking progress in what is a limited EMI assessment field. The presentation highlights key aspects of research design, analysis, and initial findings, and should be of value to those interested in EMI in HE and L2 speaking/assessment in academic contexts.

Poster 11: Towards Inclusive Language Assessment: Enhancing Equity and Diversity in Test Design and Practice

Sarvinoz Umarova, Urgench State University

Ensuring equity, diversity, and inclusion (EDI) in language assessment has become an essential priority in contemporary applied linguistics. Traditional language testing methods may inadvertently marginalize test-takers from diverse linguistic and cultural backgrounds, particularly when test content, formats, or scoring mechanisms fail to account for such variability. This study explores how EDI-oriented approaches can enhance the validity and fairness of language assessment practices across different educational contexts. Through a critical review of empirical studies and emerging frameworks in inclusive assessment, this paper highlights key factors such as test accessibility, linguistic background, disability accommodations, and sociocultural sensitivity. It further considers how inclusive design principles-such as multimodal input, flexible timing, and representative linguistic content - can help reduce construct-irrelevant variance and ensure all learners are assessed on a level playing field. The paper also discusses how test developers and educators can implement inclusive practices without compromising reliability or standardization.

Examples from classroom-based and high-stakes assessments are used to illustrate how inclusivity enhances both the learner experience and assessment outcomes. This work contributes to a growing recognition that inclusive assessment is not only an ethical imperative but also a methodological necessity for capturing learners' true language abilities.

OXFORD
TEST OF ENGLISH
Advanced

Introducing the next evolution in computer-adaptive English testing – the Oxford Test of English Advanced:

Technical Excellence

- ☑ State-of-the-art adaptive algorithm tested through extensive piloting
- ☑ SEM monitoring for precise measurement accuracy

Academic Rigour

- ☑ Developed through extensive collaboration with leading institutions
- ☑ Aligned with CEFR levels B2-C1
- ☑ Comprehensive construct coverage based on modern theories of language proficiency
- ☑ Multiple validation studies available in open-access publications

Explore our Assessment research papers and test specifications

Practical Implementation

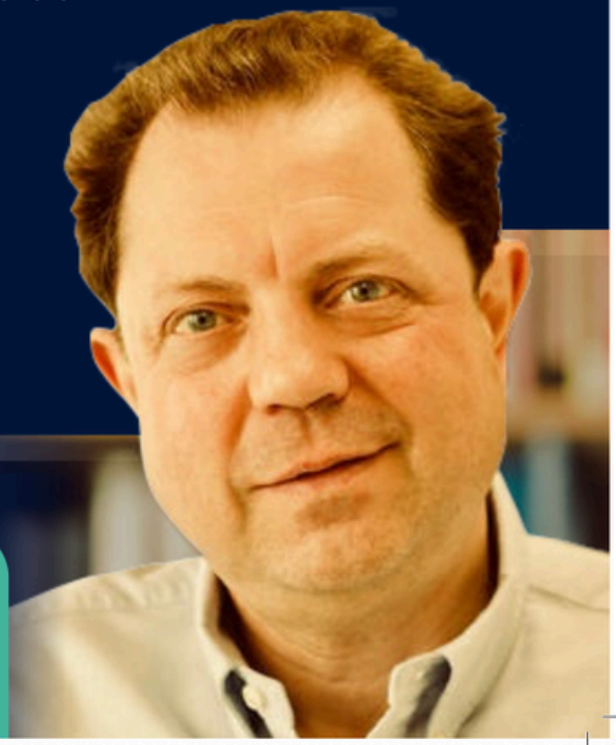
- ☑ Flexible administration options for institutional needs
- ☑ Comprehensive technical support for administrators

Research Opportunities

- ☑ Rich data for validation studies
- ☑ Collaboration possibilities for academic researchers

“ The academic orientation of the Oxford Test of English Advanced and the way its activities combine skills will help ensure that international students are fully able to engage with our courses.”

Prof Paul C. Irwin Crookes
Director of Graduate Studies, Oxford School of Global and Area Studies
University of Oxford



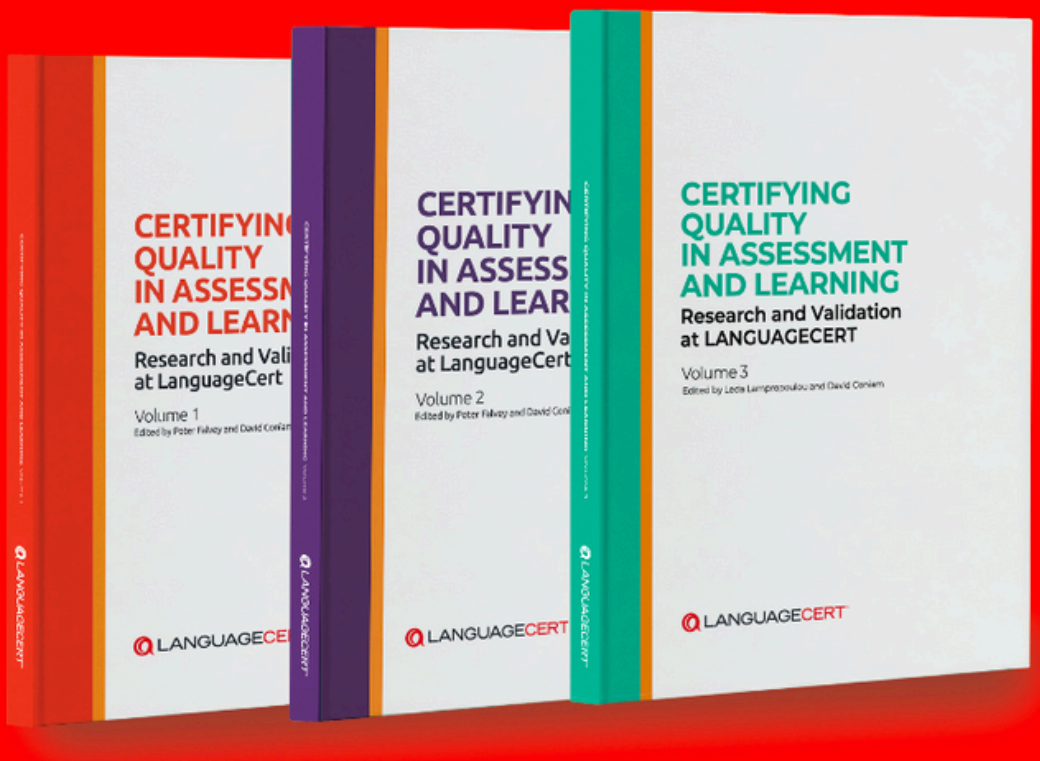
LANGUAGECERT®

Discover our LANGUAGECERT Research & Validation Series

Our English tests

- are trusted by governments, universities and employers globally
- backed by rigorous research, continuous quality assurance, and global benchmarking.

Find out how we ensure **validity**, **reliability**, and **fairness** in our English tests.



Download your copies at languagecert.org

PeopleCert
Dream it, do it.
25 YEARS

Kaplan Test of English

A new generation of English language assessment

Kaplan Test of English (KTE) is a computer-adaptive proficiency test designed for higher education admissions.

Designed and validated by Kaplan's assessment experts, KTE delivers precise and secure measurement of Listening, Reading, Writing, and Speaking.

Combining advanced psychometric modelling with rigorous linguistic construct design, KTE provides a CEFR-aligned measurement of academic English proficiency.

Hybrid automatic-human scoring ensures efficiency without compromising construct validity or fairness, while adaptive delivery enhances precision at every ability level.

Key features

- ☆ **Single test** — measures Listening, Reading, Writing and Speaking in one session.
- ☆ **Worldwide availability** — accessible on-demand, no test centre needed.
- ☆ **Fast results** — certificate provided within 2-5 days and instantly verifiable.
- ☆ **Trusted globally** — accepted by over 120 higher education institutions for admissions.
- ☆ **Suitable for all levels (A1-C2)** – CEFR-alignment independently benchmarked by Ecctis.

KAPLAN TESTING SERVICES

Learn more about our test development and validation

Learn more about the independent evaluation and benchmarking of KTE

光华启迪
GUANGHUA QIDI EDUCATION

Guanghua Qidi College

「School of one」

The Trailblazer for Customized International Education

Guanghua Qidi College is a personalized academy as the subsidiary of Guanghua Education Group which has accumulated experience in international education for many years. Since the first campus was formally established in April 2015, we have been contributing to providing the students with personalized and diversified international education in the frame of the one-stop service system.

For the purpose of cultivating and enhancing the academic competitiveness and cultural adaptability of students who aim in enrolling in the world's prestigious universities in the future, every mentor from Guanghua Qidi has been adhering to the belief of 'teaching students according to their aptitude'. Since we introduced the concept of 'School of One' into China for the first time, we have been making efforts to apply this unique concept into our personalized pedagogy to prepare high school students who have the aspiration and necessary skills to become lifelong learners and leaders.

Achievements(since 2016)

Offers(G5+HKU)

Oxford: 9
Cambridge: 41
IC: 132
LSE: 34
UCL: 538
HKU: 63

1000+ Graduates

Since 2016, Guanghua Qidi has had over 1,000 graduates.

Lancaster University has an excellent international reputation for postgraduate teaching and research in language testing and assessment

Lancaster University

Offered by our expert team:

- **Master's in Language Testing (Online)** – two years part-time (4 modules and dissertation)
- **Postgraduate Diploma in Language Testing (Online)** – two years part-time (4 modules)
- **Postgraduate Certificate in Language Testing (Online)** – one year part-time (2 modules)
- **Short courses** on topics in Language Testing – 15 weeks part-time online
  ideal for professional development or personal interest
- **PhD in Linguistics** – by research or thesis and coursework, full- or part-time, on campus or online

We also provide consultancy services on language test development and validation.

Enquiries: socialsciencesteaching@lancaster.ac.uk

www.lancaster.ac.uk/social-sciences/linguistics-and-english-language/

Linguistics at Lancaster University is ranked 3rd in the world – QS World Subject Rankings 2025